Machine Learning (2019) 108:2087–2111
https://doi.org/10.1007/s10994-019-05814-1

# Speculate-correct error bounds for *k*-nearest neighbor classifiers

**Eric Bax**[1] · **Lingjie Weng**[2] · **Xu Tian**[3]

Received: 22 September 2017 / Revised: 18 May 2018 / Accepted: 29 May 2019 / Published online: 18 June 2019


**Abstract**
We introduce the speculate-correct method to derive error bounds for local classifiers. Using it, we show that $k$-nearest neighbor classifiers, in spite of their famously fractured decision boundaries, have exponential error bounds with $\mathrm{O}\left(\sqrt{(k+\ln n)/n}\right)$ range around an estimate of generalization error for $n$ in-sample examples.

**Keywords** Nearest neighbors · Error bounds · Generalization

## 1 Introduction

Local classifiers use only a small subset of their examples to classify each input. The best-known local classifier is the nearest neighbor classifier. To classify an example, a $k$-nearest neighbor ($k$-nn) classifier uses a majority vote over the $k$ in-sample examples closest to the example. Deriving error bounds for $k$-nn classifiers is a challenge, because they can have extremely fractured decision boundaries, making approaches based on hypothesis class size ineffective. For general information on $k$-nn classifiers, see the books by Devroye et al. (1996), Duda et al. (2001) and Hastie et al. (2009).

The error bounds in this paper are probably approximately correct (PAC) bounds, consisting of a range of error rates and an upper bound on the probability that the out-of-sample error rate is outside the range. An effective PAC bound has a small range and a small bound failure probability. PAC error bounds include bounds based on Vapnik–Chervonenkis (VC)

Editor: Tapio Elomaa.

✉ Eric Bax
baxhome@yahoo.com

Lingjie Weng
lingjieweng@gmail.com

Xu Tian
tianxu03@gmail.com

[1] Verizon, Playa Vista, USA

[2] LinkedIn, Mountain View, USA

[3] Sorin Capital Management, Stamford, USA

dimension (Vapnik and Chervonenkis 1971), bounds for concept learning by Valiant (1984), compression-based bounds by Littlestone and Warmuth (1986), Floyd and Warmuth (1995), Blum and Langford (2003), and Bax (2008), and bounds based on worst likely assignments (Bax and Callejas 2008). Langford (2005) gives an overview and comparison of some types of PAC bounds.

Exponential error bounds have range proportional to $\sqrt{\ln(1/\delta)}$ as bound failure probability $\delta \to 0$. Devroye et al. (1996) (page 414) give $k$-nn classifier error bounds that are non-exponential (range proportional to $\sqrt{\frac{1}{\delta}}$ as $\delta \to 0$) and have range $\mathrm{O}\left((\sqrt{k}/n)^{1/2}\right)$. They state: "Exponential upper bounds ... are typically much harder to obtain." Then they present an exponential bound by Devroye and Wagner (1979) with range $\mathrm{O}\left((k/n)^{1/3}\right)$ (Devroye et al. (1996) p. 415, Theorem 24.5). A more recent exponential bound has expected (but not guaranteed) error bound range $\mathrm{O}\left((k/n)^{2/5}\right)$ (Bax 2012).

The great conundrum of classifier validation is that we want to use data that are independent of the classifier to estimate its error rate, but we also want to use all available data for the classifier. At each step, speculate-correct assumes that this problem does not exist, at least for some of the in-sample data. In subsequent steps, it corrects for its sometimes-false earlier assumptions. As it does this, the number of corrections grows, but the size of each correction shrinks.

To illustrate, for some value $m \leq \frac{1}{2}(n-k)$, let $V_1$ be the first $m$ and $V_2$ be the second $m$ in-sample examples. (Call $V_1$ and $V_2$ *validation subsets*.) Let $g$ be the full classifier; our goal is to bound its error rate: $Pr\{\overline{g}\}$. (Use $Pr\{\}$ to indicate probability over out-of-sample examples, and use a bar on top to indicate classifier error.) Let $g_{-S}$ be the classifier formed by withholding the data sets indexed by $S$. For example, $g_{-\{1\}}$ is all in-sample examples except those in $V_1$. Then the speculate-correct process is:

1. Speculate that withholding $V_1$ does not affect classification: $\forall x : g = g_{-\{1\}}$. Compute $Pr_{V_1}\left\{\overline{g_{-\{1\}}}\right\}$ as our initial estimate of $Pr\{\overline{g}\}$. (Use $Pr_{V_i}\{\}$ to indicate empirical rate over examples in $V_i$—also called an empirical mean.) Split the probabilities by whether the speculation holds:

$$Pr\{\overline{g}\} = Pr\left\{g = g_{-\{1\}} \wedge \overline{g}\right\} + Pr\left\{g \neq g_{-\{1\}} \wedge \overline{g}\right\}, \tag{1}$$

and

$$Pr\left\{\overline{g_{-\{1\}}}\right\} = Pr\left\{g = g_{-\{1\}} \wedge \overline{g_{-\{1\}}}\right\} + Pr\left\{g \neq g_{-\{1\}} \wedge \overline{g_{-\{1\}}}\right\}. \tag{2}$$

The RHS first terms are equal:

$$Pr\left\{g = g_{-\{1\}} \wedge \overline{g}\right\} = Pr\left\{g = g_{-\{1\}} \wedge \overline{g_{-\{1\}}}\right\}, \tag{3}$$

since

$$(g = g_{-\{1\}}) \implies (\overline{g} = \overline{g_{-\{1\}}}). \tag{4}$$

So the bias in estimating $Pr\{\overline{g}\}$ by $Pr_{V_1}\left\{\overline{g_{-\{1\}}}\right\}$ is the other two RHS terms:

$$Pr\{\overline{g}\} - Pr\left\{\overline{g_{-\{1\}}}\right\} = Pr\left\{g \neq g_{-\{1\}} \wedge \overline{g}\right\} - Pr\left\{g \neq g_{-\{1\}} \wedge \overline{g_{-\{1\}}}\right\}. \tag{5}$$

Note that $g \neq g_{-\{1\}}$ (failure of our speculation) is a condition in both bias terms. Also, the bias is in a range bounded by the probability that speculation fails:

$$\left[-Pr\left\{g \neq g_{-\{1\}}\right\}, Pr\left\{g \neq g_{-\{1\}}\right\}\right]. \tag{6}$$

2. To correct for bias due to failure of speculation in Step 1, now speculate that $\forall x : g = g_{-\{2\}}$ and $g_{-\{1\}} = g_{-\{1,2\}}$, in other words, that removing $V_2$ does not alter classifications. Then use empirical means over $V_2$ to correct bias from Step 1:

(a) Estimate $Pr\left\{g \neq g_{-\{1\}} \wedge \overline{g}\right\}$ by $Pr_{V_2}\left\{g_{-\{2\}} \neq g_{-\{1,2\}} \wedge \overline{g_{-\{2\}}}\right\}$.
(b) Estimate $-Pr\left\{g \neq g_{-\{1\}} \wedge \overline{g_{-\{1\}}}\right\}$ by $-Pr_{V_2}\left\{g_{-\{2\}} \neq g_{-\{1,2\}} \wedge \overline{g_{-\{1,2\}}}\right\}$.

Consider Estimate 2a. Split the probabilities by whether the new speculation holds:

$$Pr\left\{g \neq g_{-\{1\}} \wedge \overline{g}\right\} \tag{7}$$
$$= Pr\left\{g \neq g_{-\{1\}} \wedge (g = g_{-\{2\}} \wedge g_{-\{1\}} = g_{-\{1,2\}}) \wedge \overline{g}\right\} \tag{8}$$
$$+ Pr\left\{g \neq g_{-\{1\}} \wedge \neg(g = g_{-\{2\}} \wedge g_{-\{1\}} = g_{-\{1,2\}}) \wedge \overline{g}\right\}, \tag{9}$$

and

$$Pr\left\{g_{-\{2\}} \neq g_{-\{1,2\}} \wedge \overline{g_{-\{2\}}}\right\} \tag{10}$$
$$= Pr\left\{g_{-\{2\}} \neq g_{-\{1,2\}} \wedge (g = g_{-\{2\}} \wedge g_{-\{1\}} = g_{-\{1,2\}}) \wedge \overline{g_{-\{2\}}}\right\} \tag{11}$$
$$+ Pr\left\{g_{-\{2\}} \neq g_{-\{1,2\}} \wedge \neg(g = g_{-\{2\}} \wedge g_{-\{1\}} = g_{-\{1,2\}}) \wedge \overline{g_{-\{2\}}}\right\}. \tag{12}$$

The RHS first terms are equal, because $g = g_{-\{2\}} \wedge g_{-\{1\}} = g_{-\{1,2\}}$ implies $(g \neq g_{-\{1\}}) = (g_{-\{2\}} \neq g_{-\{1,2\}})$ and $\overline{g} = \overline{g_{-\{2\}}}$. The RHS second terms contribute bias:

$$Pr\left\{g \neq g_{-\{1\}} \wedge \overline{g}\right\} - Pr\left\{g_{-\{2\}} \neq g_{-\{1,2\}} \wedge \overline{g_{-\{2\}}}\right\} \tag{13}$$
$$= Pr\left\{g \neq g_{-\{1\}} \wedge \neg(g = g_{-\{2\}} \wedge g_{-\{1\}} = g_{-\{1,2\}}) \wedge \overline{g}\right\} \tag{14}$$
$$- Pr\left\{g_{-\{2\}} \neq g_{-\{1,2\}} \wedge \neg(g = g_{-\{2\}} \wedge g_{-\{1\}} = g_{-\{1,2\}}) \wedge \overline{g_{-\{2\}}}\right\}. \tag{15}$$

A similar analysis of Estimate 2b yields two more bias terms. All four bias terms have failure of both the first and second speculations in their conditions. So if we estimate $Pr\{\overline{g}\}$ by

$$Pr_{V_1}\left\{\overline{g_{-\{1\}}}\right\} \tag{16}$$
$$+ Pr_{V_2}\left\{g_{-\{2\}} \neq g_{-\{1,2\}} \wedge \overline{g_{-\{2\}}}\right\} - Pr_{V_2}\left\{g_{-\{2\}} \neq g_{-\{1,2\}} \wedge \overline{g_{-\{1,2\}}}\right\}, \tag{17}$$

the estimate has four bias terms, and lies in a range determined by the probability that both speculations fail:

$$[-2Pr\left\{g \neq g_{-\{1\}} \wedge \neg(g = g_{-\{2\}} \wedge g_{-\{1\}} = g_{-\{1,2\}})\right\}, \tag{18}$$
$$2Pr\left\{g \neq g_{-\{1\}} \wedge \neg(g = g_{-\{2\}} \wedge g_{-\{1\}} = g_{-\{1,2\}})\right\}]. \tag{19}$$

Continuing this for $r$ steps, with $r$ validation subsets $V_1, \ldots, V_r$, produces a sum of $2^r - 1$ estimates. All remaining bias depends on simultaneous failure of $r$ speculations, but there are $2^r$ bias terms. For $k$-nn, speculation can only fail for Step $i$ if $V_i$ has a nearer neighbor to $x$ than its $kth$ nearest neighbor among the in-sample examples not in any validation subset. So the bias is at most $2^r$ times the probability that $x$ has a nearer neighbor in every validation subset than the $kth$ nearest neighbor among the other in-sample examples.

To produce effective error bounds, we must use validation subsets small enough to make the probability of $r$ simultaneous speculation failures small, and yet large enough that the sum of $2^r - 1$ estimates is likely to have a small deviation from the sum that it estimates. (Using Hoeffding bounds (Hoeffding 1963), the range for the difference between each estimate $Pr_{V_i}\{\}$ and its corresponding out-of-sample probability $Pr\{\}$ is $\mathrm{O}\left(\frac{1}{\sqrt{|V_i|}}\right)$.) We show that an appropriate choice of validation subset size gives error bound range:

$$\mathrm{O}\left(n^{-\frac{r}{2r+1}}\sqrt{k+r}\right), \tag{20}$$

and, for a choice of $r$ based on $n$, the range is

$$\mathrm{O}\left(\sqrt{(k+\ln n)/n}\right). \tag{21}$$

The next section formally introduces the speculate-correct method to produce error bounds for local classifiers. Section 3 applies the method to $k$-nn classifiers. Section 4 shows how to compute the bounds. Section 5 shows how effective the bounds are for some actual classifiers. Section 6 concludes with potential directions for future work.

## 2 Speculate-Correct

Let $F$ be the full set of $n$ in-sample examples $(x, y)$, drawn i.i.d. from a joint input-output distribution $D$. Inputs $x$ are drawn from an arbitrary domain, and outputs $y$ are drawn from $\{0, 1\}$ (binary classification). Assume there is some ordering of the examples in $F$, so that we may refer to examples 1 to $n$ in $F$, treating $F$ as a sequence.

Select $r > 0$ and $m > 0$ such that $rm \leq n - k$. For each $i \in 1, \ldots, r$, let validation subset $V_i$ be the $i$th subset of $m$ examples in $F$. For example, if $r = 2$ and $m = 1000$, then $V_1$ is the first thousand examples in $F$ and $V_2$ is the second thousand. Let validation set $V = V_1 \cup \ldots \cup V_r$. For convenience, define $R \equiv \{1, \ldots, r\}$. For $S \subseteq R$, let $V_S$ be the union of validation subsets indexed by $S$.

Our PAC error bounds have probability of bound failure over draws of $F$. Let the subscript $F \sim D^n$ denote a probability or expectation over draws of $F$. We use no subscript for probabilities or expectations over out-of-sample examples $(x, y) \sim D$. For example,

$$p^* \equiv Pr\{\overline{g}\} \tag{22}$$

denotes the out-of-sample error rate of $g$, and it is the quantity we wish to bound. (It is sometimes called the conditional error rate, because it is the error rate conditioned on a set of in-sample examples $F$ rather than the expected error rate over draws of $F$.)

Let $A_i = \{1, \ldots, i\}$. Let $a_1, \ldots, a_r$ be any series of conditions such that

$$a_i(x) \implies \forall S \subseteq A_{i-1} : g_{-(S\cup\{i\})}(x) = g_{-S}(x), \tag{23}$$

i.e., $a_i(x)$ implies that for any classifier formed by withholding any subset of $\{V_1, \ldots, V_{i-1}\}$, withholding $V_i$ too does not alter the classification of $x$.

Let $b_i = \neg a_1 \wedge \ldots \wedge \neg a_i$. Define $b_0$ to be true. The following theorem generalizes the speculate-correct formula for $r = 2$ that we developed in the previous section.

**Theorem 1**

$$\forall r \geq 0 : p^* = \sum_{i=1}^{r} \sum_{S\subseteq A_{i-1}} (-1)^{|S|} Pr\left\{b_{i-1} \wedge \overline{g_{-(S\cup\{i\})}}\right\} + \sum_{S\subseteq R} (-1)^{|S|} Pr\left\{b_r \wedge \overline{g_{-S}}\right\}. \tag{24}$$

***Proof*** Use induction. The base case is $r = 0$:

$$\sum_{S\subseteq\emptyset} (-1)^{|S|} Pr\{b_0 \wedge \overline{g_{-S}}\} = Pr\{b_0 \wedge \overline{g_{-\emptyset}}\} = Pr\{\overline{g}\} = p^*. \tag{25}$$

Next, to show that the result for $r$:

$$p^* = \sum_{i=1}^{r} \sum_{S\subseteq A_{i-1}} (-1)^{|S|} Pr\left\{b_{i-1} \wedge \overline{g_{-(S\cup\{i\})}}\right\} + \sum_{S\subseteq R} (-1)^{|S|} Pr\left\{b_r \wedge \overline{g_{-S}}\right\}, \tag{26}$$

implies the result for $r+1$:

$$p^* = \sum_{i=1}^{r+1} \sum_{S \subseteq A_{i-1}} (-1)^{|S|} Pr\left\{b_{i-1} \wedge \overline{g_{-(S \cup \{i\})}}\right\} + \sum_{S \subseteq A_{r+1}} (-1)^{|S|} Pr\left\{b_{r+1} \wedge \overline{g_{-S}}\right\}, \quad (27)$$

subtract the result for $r$ from the result for $r+1$. The difference is

$$\sum_{S \subseteq A_r} (-1)^{|S|} Pr\left\{b_r \wedge \overline{g_{-(S \cup \{r+1\})}}\right\} + \sum_{S \subseteq A_{r+1}} (-1)^{|S|} Pr\left\{b_{r+1} \wedge \overline{g_{-S}}\right\} \quad (28)$$

$$- \sum_{S \subseteq R} (-1)^{|S|} Pr\left\{b_r \wedge \overline{g_{-S}}\right\}. \quad (29)$$

We will show that this difference is zero.

Since $A_r = R$, the first and third sums are over the same indices, so combine them:

$$= \sum_{S \subseteq A_r} (-1)^{|S|} \left(Pr\left\{b_r \wedge \overline{g_{-(S \cup \{r+1\})}}\right\} - Pr\left\{b_r \wedge \overline{g_{-S}}\right\}\right) \quad (30)$$

$$+ \sum_{S \subseteq A_{r+1}} (-1)^{|S|} Pr\left\{b_{r+1} \wedge \overline{g_{-S}}\right\}. \quad (31)$$

Expand the first sum's probabilities around $a_{r+1}$ values:

$$Pr\left\{b_r \wedge \overline{g_{-(S \cup \{r+1\})}}\right\} - Pr\left\{b_r \wedge \overline{g_{-S}}\right\} \quad (32)$$

$$= Pr\left\{b_r \wedge a_{r+1} \wedge \overline{g_{-(S \cup \{r+1\})}}\right\} + Pr\left\{b_r \wedge \neg a_{r+1} \wedge \overline{g_{-(S \cup \{r+1\})}}\right\} \quad (33)$$

$$- Pr\left\{b_r \wedge a_{r+1} \wedge \overline{g_{-S}}\right\} - Pr\left\{b_r \wedge \neg a_{r+1} \wedge \overline{g_{-S}}\right\}. \quad (34)$$

The first and third terms cancel, because $a_{r+1} \implies g_{-(S \cup \{r+1\})} = g_{-S}$. The other terms have $b_{r+1}$, since $b_r \wedge \neg a_{r+1} = b_{r+1}$. So the difference is:

$$= \sum_{S \subseteq A_r} (-1)^{|S|} \left(Pr\left\{b_{r+1} \wedge \overline{g_{-(S \cup \{r+1\})}}\right\} - Pr\left\{b_{r+1} \wedge \overline{g_{-S}}\right\}\right) \quad (35)$$

$$+ \sum_{S \subseteq A_{r+1}} (-1)^{|S|} Pr\left\{b_{r+1} \wedge \overline{g_{-S}}\right\}. \quad (36)$$

The first sum cancels the second: for each $S$ in the first sum, the first term cancels the term for $S \cup \{r+1\}$ in the second sum, and the second term cancels the term for $S$ in the second sum. □

The formulation of the error rate in Theorem 1 is useful because the examples in each validation subset $V_i$ are independent of the conditions in term $i$ in the first sum. So the rates of the conditions over the validation subsets are unbiased estimates of the probabilities of those conditions over out-of-sample examples. There are no such validation data for the second sum. Instead of estimating the second sum, our error bounds bound each of its terms by $Pr\{b_r\}$. We select validation subset sizes to mediate a tradeoff: large validation subsets give tight bounds on terms in the first sum, but small validation subsets make $Pr\{b_r\}$ small.

## 3 Error bounds for *k*-NN classifiers

Before introducing $k$-nn error bounds, we need a brief aside about tie-breaking. Assume $k$ is odd and assume binary classification, so there are no ties in voting. To break ties over which

in-sample examples are nearest neighbors and hence vote, use the method from Devroye and Wagner (1979): assign each example $i$ in $F$ a real value $Z_i$ drawn uniformly at random from [0, 1] and do the same for each other draw $x$ from the input space to give it a value $Z$. If the distance from example $i$ in $F$ to an $x$ is the same as the distance from example $j$ in $F$ to $x$, then declare $i$ to be the closer example if $|Z_i - Z| < |Z_j - Z|$ or if $|Z_i - Z| = |Z_j - Z|$ and $i < j$. Otherwise declare example $j$ to be the closer example. This method returns the same ranking of distances to examples in $F$ for the same input $x$ every time the distances are measured, and it uses position within $F$ to break a tie with probability zero.

Now apply the speculate-correct concept to $k$-nn:

**Corollary 2** *Let $a_i(x)$ be the condition that $V_i$ does not have an example closer to $x$ than the $k$th nearest neighbor to $x$ in $F - V$. Let*

$$\forall 1 \leq i \leq r : f_i(x, y) = I(b_{i-1}) \sum_{S \subseteq A_{i-1}} (-1)^{|S|} I\left(\overline{g_{-(S \cup \{i\})}}\right) \tag{37}$$

*and*

$$f_{r+1}(x, y) = I(b_r) \sum_{S \subseteq A_r} (-1)^{|S|} I\left(\overline{g_{-S}}\right), \tag{38}$$

*where $I()$ is the indicator function: one if its argument is true and zero otherwise. Then*

$$p^* = \sum_{i=1}^{r+1} E\{f_i\}. \tag{39}$$

***Proof*** Our $a_i$ for $k$-nn meet the conditions of Theorem 1. □

Next, we show that $p^*$ is the average of the RHS of Eq. 39 from Corollary 2 over all permutations of the in-sample examples. Permuting the examples places different examples into the validation subsets because the $ith$ validation subset is the $ith$ $m$ examples. We will use permutations to ensure that $b_1, \ldots, b_r$ are rare enough to provide small error bound ranges.

Without permutations, even $b_r$ may not be rare. For example, in-sample examples $m, 2m, \ldots, rm$ may all be close to much of the input distribution, and the other in-sample examples may be far. Without permutations, we can develop a bound, but we can only show that it has a small error bound range in expectation. Permutations guarantee that the expectation is realized. In the next section, we show how to compute permutation-based bounds efficiently.

**Lemma 3** *Let $P$ be the set of permutations of $1, \ldots, n$. For each $\sigma \in P$, let $\sigma F$ be $F$ permuted according to $\sigma$: example $j$ of $\sigma F$ is the example of $F$ indexed by element $j$ of $\sigma$. Let $f_{i,\sigma}$ be $f_i$, but with $F$ replaced by $\sigma F$, so that for $i \in R$, $V_i$ consists of the $ith$ $m$ examples in $\sigma F$. Then*

$$p^* = E_{\sigma \in P}\left\{\sum_{i=1}^{r+1} E\left\{f_{i,\sigma}\right\}\right\}. \tag{40}$$

***Proof*** Corollary 2 holds for each partition of $F$ into $r$ size-$m$ subsets $V_1, \ldots, V_r$ and $F - V$. Each permutation of $F$ uses one of these partitions to define $f_{i,\sigma}$. So the outer expectation is over quantities that are each $p^*$. □

We will use two more lemmas to form a bound based on permutations:

**Lemma 4** *For any set of permutations $P'$,*

$$\forall x, i > 1 : \left|E_{\sigma\in P'}\left\{f_{i,\sigma}\right\}\right| \leq 2^{i-2} Pr_{\sigma\in P'}\left\{b_{i-1}|\sigma\right\}. \tag{41}$$

*For $i = 1$,*

$$\forall x : E_{\sigma\in P'}\left\{f_{i,\sigma}\right\} \in [0, 1]. \tag{42}$$

***Proof*** For $i > 1$,

$$\left|E_{\sigma\in P'}\left\{f_{i,\sigma}\right\}\right| \leq E_{\sigma\in P'}\left\{\left|f_{i,\sigma}\right|\right\} \tag{43}$$

Since $f_{i,\sigma}$ is a sum of $2^{i-1}$ terms, with half 0 or 1 and half 0 or $-1$,

$$\left|f_{i,\sigma}\right| \leq 2^{i-2} I(b_{i-1}|\sigma). \tag{44}$$

So

$$E_{\sigma\in P'}\left\{\left|f_{i,\sigma}\right|\right\} \leq 2^{i-2} E_{\sigma\in P'}\left\{I(b_{i-1}|\sigma)\right\} = 2^{i-2} Pr_{\sigma\in P'}\left\{b_{i-1}|\sigma\right\}. \tag{45}$$

For $i = 1$, note that $f_{i,\sigma}$ has a single term, with value zero or one. □

**Lemma 5** *Let $P'$ be a set of permutations such that for each $\sigma \in P'$, positions $1, \ldots, im, rm + 1, \ldots, n$ of the permutations in $P'$ contain all permutations of entries in those positions in $\sigma$, equally many times. Then*

$$Pr_{\sigma\in P'}\left\{b_{i-1}|\sigma\right\} = \sum_{h=0}^{i-1}(-1)^h \binom{i-1}{h} \prod_{j=0}^{k-1} \frac{n - rm - j}{n - rm + hm - j}. \tag{46}$$

***Proof*** The LHS is the probability that a random permutation in $P'$ places at least one example in each of $V_1, \ldots, V_{i-1}$ that is closer to $(x, y)$ than the $kth$ closest example to $(x, y)$ in $F - V$. Since determining positions in a random draw from $P'$ is equivalent to drawing positions at random without replacement, the LHS is the probability of drawing at least one element from each set $\{1, \ldots, m\}, \ldots, \{(i-1)m + 1, \ldots, im\}$ before drawing $k$ elements from $\{rm + 1, \ldots, n\}$.

The probability of drawing $k$ elements from $\{rm + 1, \ldots, n\}$ before drawing any from one specific set in $\{1, \ldots, m\}, \ldots, \{(i-2)m + 1, \ldots, (i-1)m\}$ is

$$\left(\frac{n - rm}{n - rm + m}\right)\left(\frac{n - rm - 1}{n - rm + m - 1}\right)\cdots\left(\frac{n - rm - (k-1)}{n - rm + m - (k-1)}\right). \tag{47}$$

Similarly, the probability of drawing $k$ elements from $\{rm + 1, \ldots, n\}$ before drawing any elements from any specific $h$ of the $i - 1$ sets in $\{1, \ldots, m\}, \ldots, \{(i-2)m+1, \ldots, (i-1)m\}$ is

$$\left(\frac{n - rm}{n - rm + hm}\right)\left(\frac{n - rm - 1}{n - rm + hm - 1}\right)\cdots\left(\frac{n - rm - (k-1)}{n - rm + hm - (k-1)}\right). \tag{48}$$

So, by inclusion and exclusion, the probability of drawing at least one element from every set in $\{1,\ldots,m\},\ldots\{(i-2)m+1,\ldots,(i-1)m\}$ before drawing $k$ examples from $\{rm+1,\ldots,n\}$ is:

$$\sum_{h=0}^{i-1}(-1)^h\binom{i-1}{h}\prod_{j=0}^{k-1}\frac{n-rm-j}{n-rm+hm-j}. \tag{49}$$

□

Note that $i=r+1$ and $P'=P$ gives a result for $Pr_{\sigma\in P}\{b_r|\sigma\}$.

We need another lemma for the bound. This one is about averaging bounds on differences between empirical means and actual means.

**Lemma 6** *Suppose there is a finite set of distributions, each with range size (difference between maximum and minimum values in support) at most $s$. For each distribution $i$ in the set, let $\mu_i$ be the mean, and let $\hat{\mu}_i$ be an empirical mean based on $m$ i.i.d. samples from distribution $i$. Let $E_i\{\}$ denote expectation over the distributions. Then*

$$\forall c>0,\delta>0: Pr\left\{\left|E_i\{\mu_i\}-E_i\left\{\hat{\mu}_i\right\}\right|\geq\epsilon\right\}\leq\delta, \tag{50}$$

*where*

$$\epsilon=\frac{s}{\sqrt{2m}}\left[\sqrt{\ln\frac{2}{\delta}}\left(\frac{e^c}{e^c-1}\right)+\sqrt{c+1}\left(\frac{e^c}{e^c-1}\right)^2+1\right]. \tag{51}$$

*Specifically, for $c=3$,*

$$\epsilon\leq\frac{s}{\sqrt{2m}}\left(1.06\sqrt{\ln\frac{2}{\delta}}+3.22\right). \tag{52}$$

***Proof*** For the proof, refer to Bax and Kooti (2016), page 3, Inequalities 8 and 9. We use $\frac{2}{\delta}$ in place of the $\frac{1}{\delta}$ found there, because we have two-sided bounds. □

This lemma offers bounds on differences between averages of means and averages of estimates that are similar to the Hoeffding bound (Hoeffding 1963) on the difference between a single mean and estimate:

$$\forall\delta>0: Pr\left\{|\mu-\hat{\mu}|\geq\frac{s}{\sqrt{2m}}\sqrt{\ln\frac{2}{\delta}}\right\}\leq\delta. \tag{53}$$

Now separate the RHS of Eq. 40 into terms with $i\in R$ and a term with $i=r+1$:

$$p^*=\left(\sum_{i=1}^{r}E_{\sigma\in P}\left\{E\left\{f_{i,\sigma}\right\}\right\}\right)+E_{\sigma\in P}\left\{E\left\{f_{r+1,\sigma}\right\}\right\} \tag{54}$$

$$=p_I+p_{II}. \tag{55}$$

To develop an error bound, we will estimate $p_I$ with empirical means over samples and bound $p_{II}$. For each $\sigma\in P$ and $i\in R$, the examples in $V_i|\sigma$ are independent of the function $f_{i,\sigma}$, so we can use empirical means over $(x,y)\in V_i|\sigma$ to estimate means over $(x,y)\sim D$. First, we rewrite $p_I$ in a form that allows estimation by empirical means over permutations, in the following lemma.

**Lemma 7** *$\forall i \in R$, let $|V_i| = m$. Let $M$ be the set of size-$m$ subsets of $F$: $M = \{Q | Q \subseteq F \wedge |Q| = m\}$. Let $P(Q, i)$ be the set of permutations of $1, \ldots, n$ that have set $Q$ as validation subset $V_i$ in $\sigma$ $F$: $P(Q, i) = \{\sigma | (V_i | \sigma) = Q\}$. Then*

$$p_I = E_{Q \in M} \left\{ \sum_{i=1}^{r} E_{\sigma \in P(Q,i)} \left\{ E \left\{ f_{i,\sigma} \right\} \right\} \right\} \tag{56}$$

***Proof*** Compare the definition of $p_I$ (the first term on the RHS of Eq. 54) to Eq. 56. The definition averages over permutations in $P$ and $i \in R$. In Eq. 56, the expectation over $Q \in M$, $i \in R$, and $P(Q, i)$ covers all permutations $P$ and $i \in R$, each with equal frequency. □

Now we develop an error bound.

**Theorem 8** *Let*

$$\hat{p}_Q = E_{(x,y) \in Q} \left\{ \sum_{i=1}^{r} E_{\sigma \in P(Q,i)} \left\{ f_{i,\sigma}(x, y) \right\} \right\} \tag{57}$$

*and*

$$\hat{p}_I = E_{Q \in M} \left\{ \hat{p}_Q \right\}. \tag{58}$$

*Then*

$$\forall \delta > 0 : Pr_{F \sim D^n} \left\{ |p^* - \hat{p}_I| \geq \epsilon_I + \epsilon_{II} \right\} \leq \delta, \tag{59}$$

*where*

$$\epsilon_I = \sum_{i=1}^{r} 2^{i-1} Pr_{\sigma \in P} \{ b_{i-1} | \sigma \} \frac{1}{\sqrt{2m}} \left( 1.06 \sqrt{\ln \frac{2}{\delta}} + 3.22 \right), \tag{60}$$

$$\epsilon_{II} = 2^{r-1} Pr_{\sigma \in P} \{ b_r | \sigma \}, \tag{61}$$

*and*

$$Pr_{\sigma \in P} \{ b_i | \sigma \} = \sum_{h=0}^{i} (-1)^h \binom{i}{h} \prod_{j=0}^{k-1} \frac{n - rm - j}{n - rm + hm - j}. \tag{62}$$

***Proof*** Note that

$$|p^* - \hat{p}_I| = |(p_I + p_{II}) - \hat{p}_I| \leq |p_I - \hat{p}_I| + |p_{II}|. \tag{63}$$

We will show that

$$\forall \delta > 0 : Pr_{F \sim D^n} \left\{ |p_I - \hat{p}_I| \geq \epsilon_I \right\} \leq \delta, \tag{64}$$

and that $|p_{II}| \leq \epsilon_{II}$. (The formula for $Pr_{\sigma \in P} \{ b_i | \sigma \}$, to specify values for $Pr_{\sigma \in P} \{ b_{i-1} | \sigma \}$ in $\epsilon_I$ and $Pr_{\sigma \in P} \{ b_r | \sigma \}$ in $\epsilon_{II}$, follows directly from Lemma 5.)

To prove Inequality 64, let

$$p_Q = E_{(x,y) \sim D} \left\{ \sum_{i=1}^{r} E_{\sigma \in P(Q,i)} \left\{ f_{i,\sigma}(x, y) \right\} \right\}. \tag{65}$$

Then

$$p_I = E_{Q \in M} \left\{ p_Q \right\}, \tag{66}$$

since this is Inequality 56 from Lemma 7, with a different order of expectations.

For each $Q \in M$, we will use each $\hat{p}_Q$ to bound each $p_Q$, using the fact the examples in $Q$ are independent of $p_Q$. First, we need to bound the range of terms in the expectations $p_Q$ and $\hat{p}_Q$:

$$\sum_{i=1}^{r} E_{\sigma \in P(Q,i)} \left\{ f_{i,\sigma} \right\}. \tag{67}$$

By Lemma 4,

$$\left| \sum_{i=1}^{r} E_{\sigma \in P(Q,i)} \left\{ f_{i,\sigma} \right\} \right| \tag{68}$$

$$\in \left[ -\sum_{i=2}^{r} 2^{i-2} Pr_{\sigma \in P(Q,i)} \left\{ b_{i-1} | \sigma \right\}, 1 + \sum_{i=2}^{r} 2^{i-2} Pr_{\sigma \in P(Q,i)} \left\{ b_{i-1} | \sigma \right\} \right]. \tag{69}$$

Also, by Lemma 5

$$Pr_{\sigma \in P(Q,i)} \left\{ b_{i-1} | \sigma \right\} = Pr_{\sigma \in P} \left\{ b_{i-1} | \sigma \right\}, \tag{70}$$

since both $P(Q, i)$ and $P$ meet the conditions for $P'$ in Lemma 5. So

$$\sum_{i=1}^{r} E_{\sigma \in P(Q,i)} \left\{ f_{i,\sigma} \right\} \tag{71}$$

$$\in \left[ -\sum_{i=2}^{r} 2^{i-2} Pr_{\sigma \in P} \left\{ b_{i-1} | \sigma \right\}, 1 + \sum_{i=2}^{r} 2^{i-2} Pr_{\sigma \in P} \left\{ b_{i-1} | \sigma \right\} \right]. \tag{72}$$

For $i = 1$, $2^{i-1} Pr_{\sigma \in P} \left\{ b_{i-1} | \sigma \right\} = 1$. So the range is at most

$$\sum_{i=1}^{r} 2^{i-1} Pr_{\sigma \in P} \left\{ b_{i-1} | \sigma \right\}. \tag{73}$$

Note that $\hat{p}_Q$ is an average of this term (Expression 71) over $|Q| = m$ i.i.d. $(x, y)$ samples that are independent of the term (since $Q = V_i | \sigma$ for $\sigma \in P(Q, i)$ and the definition of $f_{i,\sigma}$ depends only on $V_1, \ldots, V_{i-1}, F - V | \sigma$.) Since $p_I$ is the expectation of a finite set of means $p_Q$ and $\hat{p}_I$ is the expectation of corresponding empirical means $\hat{p}_Q$, we can apply Lemma 6 to prove Inequality 64, showing that $\hat{p}_I$ is an $\epsilon_I$-range estimate of $p_I$.

For $p_{II}$ and $\epsilon_{II}$, apply Lemma 4:

$$|p_{II}| \leq 2^{r-1} Pr_{\sigma \in P} \left\{ b_r | \sigma \right\} = \epsilon_{II}. \tag{74}$$

□

We will use the following lemma to prove results about the size of the error bound range $\epsilon_I + \epsilon_{II}$.

**Lemma 9**

$$Pr_{\sigma \in P} \left\{ b_i | \sigma \right\} \leq \left( \frac{e(k + i - 1)m}{n} \right)^i. \tag{75}$$

***Proof*** Define $d_i(x) | \sigma$ to be the condition that the $k + i - 1$ nearest neighbors to $x$ in $F$ include at least $i$ examples from $V_1 \cup \ldots \cup V_i | \sigma$. Condition $d_i | \sigma$ is a necessary condition for $b_i | \sigma$, so

$$\forall x : Pr_{\sigma \in P} \left\{ d_i | \sigma \right\} \geq Pr_{\sigma \in P} \left\{ b_i | \sigma \right\}. \tag{76}$$

The probability of $d_i|\sigma$ over $\sigma \in P$ is the same as the probability of drawing $k+i-1$ samples from $1, \ldots, im, rm+1, \ldots, n$ uniformly without replacement and having at least $i$ of those samples have values $im$ or less. (The samples are the indices in $\sigma F$ of the $k+i-1$ nearest neighbors to $x$ from positions $1, \ldots, im, rm+1, \ldots, n$ in $\sigma F$.) So the probability of $d_i$ is the tail of a hypergeometric distribution:

$$\forall x : Pr_{\sigma \in P}\{d_i|\sigma\} = \sum_{j=i}^{k+i-1} \frac{\binom{k+i-1}{j}\binom{n-(k+i-1)}{im-j}}{\binom{n}{rm}}. \tag{77}$$

Using a hypergeometric tail bound from Chvátal (1979) (see also Skala (2013)), this is

$$\leq \left(\frac{(k+i-1)m}{n}\right)^i \left[\left(1+\frac{1}{m-1}\right)\left(1-\frac{k+i-1}{n}\right)\right]^{(m-1)i} \tag{78}$$

$$\leq \left(\frac{(k+i-1)m}{n}\right)^i \left[\left(1+\frac{1}{m-1}\right)^{m-1}\right]^i \tag{79}$$

$$\leq \left(\frac{(k+i-1)m}{n}\right)^i e^i. \tag{80}$$

□

**Corollary 10 (of Theorem 8)**

$$\epsilon_I + \epsilon_{II} \tag{81}$$

$$\leq \left(\frac{1}{1-\frac{2e(k+r-2)m}{n}}\right)\frac{1}{\sqrt{2m}}\left(1.06\sqrt{\ln\frac{2}{\delta}}+3.22\right)+\left(\frac{2e(k+r-1)m}{n}\right)^r \tag{82}$$

***Proof*** Recall that

$$\epsilon_I = \sum_i^r 2^{i-1} Pr_{\sigma \in P}\{b_{i-1}|\sigma\}\frac{1}{\sqrt{2m}}\left(1.06\sqrt{\ln\frac{2}{\delta}}+3.22\right). \tag{83}$$

By Lemma 9,

$$\sum_i^r 2^{i-1} Pr_{\sigma \in P}\{b_{i-1}|\sigma\} \leq \sum_{i=1}^r \left(\frac{2e(k+i-2)m}{n}\right)^{i-1}. \tag{84}$$

Apply the well-known identity for a sum of powers: $1+z+z^2+\cdots+z^{r-1} = \frac{1-z^r}{1-z}$, with $z = \frac{2e(k+r-2)m}{n}$:

$$\leq \frac{1}{1-\frac{2e(k+r-2)m}{n}}. \tag{85}$$

So

$$\epsilon_I \leq \left(\frac{1}{1-\frac{2e(k+r-2)m}{n}}\right)\frac{1}{\sqrt{2m}}\left(1.06\sqrt{\ln\frac{2}{\delta}}+3.22\right). \tag{86}$$

For $\epsilon_{II}$, apply Lemma 9:

$$\epsilon_{II} = 2^{r-1} Pr_{\sigma \in P}\{b_r|\sigma\} \leq \left(\frac{2e(k+r-1)m}{n}\right)^r. \tag{87}$$

□

The following theorem and corollary are the main results for $k$-nn classifiers. The theorem allows $r$, $k$, and $\delta$ to depend on the number of in-sample examples, $n$. The corollary uses the bound from the theorem with an appropriate growth rate for $r$ as $n$ increases.

**Theorem 11**

$$\forall \delta > 0 : Pr_{F \sim D^n} \left\{ |p^* - \hat{p}_I| \leq \epsilon_r \right\} \leq \delta, \tag{88}$$

*with*

$$\epsilon_r \in O\left( n^{-\frac{r}{2r+1}} \sqrt{(k+r)} \right). \tag{89}$$

***Proof*** Let $\epsilon_r = \epsilon_I + \epsilon_{II}$ in Theorem 8. Use Corollary refgreen for $\epsilon_I + \epsilon_{II}$. Select validation subset sizes $m$ to balance $\epsilon_I$ and $\epsilon_{II}$:

$$m = \left\lceil \frac{n^{\frac{r}{r+\frac{1}{2}}}}{2e(k+r-1)} \right\rceil. \tag{90}$$

Then

$$\epsilon_{II} \leq \left( \frac{2e(k+r-1)m}{n} \right)^r \in \mathrm{O}\left( n^{-\frac{r}{2r+1}} \right), \tag{91}$$

and

$$\epsilon_I \leq n^{-\frac{r}{2r+1}} \left( \frac{1}{1 - n^{-\frac{1}{2r+1}}} \right) \sqrt{2e(k+r-1)} \frac{1}{\sqrt{2}} \left( 1.06 \sqrt{\ln \frac{2}{\delta}} + 3.22 \right). \tag{92}$$

If we allow for the possibility of $k$ and $r$ growing with $n$, then

$$\epsilon_r = \epsilon_I + \epsilon_{II} \in \mathrm{O}\left( n^{-\frac{r}{2r+1}} \sqrt{k+r} \right). \tag{93}$$

□

**Corollary 12** *For a choice of $r$ based on $n$,*

$$\forall \delta > 0 : Pr_{F \sim D^n} \left\{ |p^* - \hat{p}_I| \leq \epsilon_* \right\} \leq \delta, \tag{94}$$

*with*

$$\epsilon_* \in O\left( \sqrt{(k + \ln n)/n} \right). \tag{95}$$

***Proof*** If we set $r = \lceil \frac{1}{4}(\ln n - 2) \rceil$, then

$$n^{-\frac{r}{2r+1}} = n^{\frac{1}{4r+2}} n^{-\frac{1}{2}} \leq n^{\frac{1}{\ln n}} n^{-\frac{1}{2}} = e n^{-\frac{1}{2}}. \tag{96}$$

So

$$\epsilon_* = \epsilon_I + \epsilon_{II} \in \mathrm{O}\left( n^{-\frac{1}{2}} \sqrt{(k + \ln n)} \right). \tag{97}$$

□

An alternative proof of Corollary 12 uses a different value for $m$:

Proof (Alternative Proof of Corollary 12) Let

$$m = \left\lfloor \frac{n}{2e^2(k+r-1)} \right\rfloor. \tag{98}$$

Then

$$\epsilon_{II} \leq \left( \frac{2e(k+r-1)m}{n} \right)^r \leq \frac{1}{e^r}. \tag{99}$$

For $\epsilon_I$, note that

$$m \geq \frac{n}{2e^2(k+r-1)} - 1 = \frac{n - 2e^2(k+r-1)}{2e^2(k+r-1)}. \tag{100}$$

Substitute the RHS for $m$ in Inequality 86:

$$\epsilon_I \leq \left(\frac{1}{1-\frac{1}{e}}\right) \frac{\sqrt{2e^2(k+r-1)}}{\sqrt{n-2e^2(k+r-1)}} \frac{1}{\sqrt{2}} \left(1.06\sqrt{\ln \frac{2}{\delta}} + 3.22\right). \tag{101}$$

Let $r = \lceil \ln \sqrt{n} \rceil$. Then

$$\epsilon_{II} \leq \frac{1}{e^{\ln \sqrt{n}}} = \frac{1}{\sqrt{n}}, \tag{102}$$

and

$$\epsilon_I \in \mathrm{O}\left(\sqrt{(k+\ln n)/n}\right). \tag{103}$$

□

## 4 Computation

It would be infeasible to compute the error bounds developed in this paper directly from their definitions. Instead, we can sample the bound terms to produce a bound. In this section, we outline a sampling procedure that requires $\mathrm{O}(n(\ln n)^2)$ computation (in addition to identifying up to $k+r-1$ nearest neighbors in $F$ for each example in $F$) and produces a bound with range $\mathrm{O}\left(\sqrt{(k+\ln n)/n}\right)$.

Note that

$$\hat{p}_I = E_{\sigma \in P}\left\{r E_{i \in R}\left\{E_{(x,y) \in V_i|\sigma}\left\{f_{i,\sigma}\right\}\right\}\right\}. \tag{104}$$

(For reference, $\hat{p}_I$ is defined in Eqs. 57 and 58 of Lemma 8.) Let $P((x,y),i) = \{\sigma | (x,y) \in (V_i|\sigma))\}$. Reordering expectations,

$$\hat{p}_I = E_{(x,y) \in F}\left\{r E_{i \in R}\left\{E_{\sigma \in P((x,y),i)}\left\{f_{i,\sigma}\right\}\right\}\right\}. \tag{105}$$

Rewrite $f_{i,\sigma}$ as the expectation of its terms:

$$f_{i,\sigma} = I(b_{i-1}) 2^{i-1} E_{S \subseteq A_{i-1}}\left\{(-1)^{|S|} I(\overline{g_{-(S \cup \{i\})}}|\sigma)\right\}. \tag{106}$$

Estimate $\hat{p}_I$ as defined in the previous two equations by taking an empirical mean over $s$ random samples:

$$((x,y), i, \sigma, S), \tag{107}$$

with $(x,y)$ drawn uniformly at random from $F$, $i$ uniformly at random from $R$, $\sigma$ uniformly at random from $P((x,y),i)$, and $S$ uniformly at random from the power set of $A_{i-1}$. Each sample value is

$$r I(b_{i-1}) 2^{i-1} (-1)^{|S|} I(\overline{g_{-(S \cup \{i\})}}|\sigma). \tag{108}$$

Let $p'_I$ be the empirical mean of these samples.

Computing values for samples in $p'_I$ need not involve drawing complete permutations $\sigma$. Instead, randomly determine set membership in $F - V|\sigma$, $V_i|\sigma$, …, or $V_r|\sigma$ for neighbors of the sample $(x,y)$ and tabulate votes to determine $I(\overline{g_{-(S \cup \{i\})}}|\sigma)$, proceeding one neighbor at a time until the $kth$ neighbor from $F - V|\sigma$ is identified, as follows.

Let $N_0(x) = (x,y)$. Let $N_j(x)$ be $(x,y)$ and the $j$ nearest neighbors to $(x,y)$ in $F$. At each step, let $f = |(F - V|\sigma) \cap N_j(x)|$. For each $i \in R$, let $v_i = |V_i \cap N_j(x)|$. Let $b$ be

the number of voting neighbors ($b$ for "ballots") among the $j$ nearest neighbors to $(x, y)$: $b = |((F - V_i) - \cup_{h \in S} V_h|\sigma) \cap N_j(x)|$. Let $d$ be the number of those voters that have different labels than $y$.

Initially, $j = 0$, $f = 0$, $v_i = 1, \forall h \neq i : v_h = 0$, $b = 0$, and $d = 0$. Then, for each $j$ starting with $j = 1$, select a set for the $jth$ nearest neighbor at random and increment its counter:

$$\begin{array}{ll} F - V|\sigma \text{ with probability } \frac{n-f-\sum_{h \in R}(m-v_h)}{n-j} : & f := f + 1 \\ V_1|\sigma \text{ with probability } \frac{m-v_1}{n-j} : & v_1 := v_1 + 1 \\ \vdots & \vdots \\ V_r|\sigma \text{ with probability } \frac{m-v_r}{n-j} : & v_r := v_r + 1 \end{array} \tag{109}$$

If $b < k$ (fewer than $k$ votes cast) and the set is $F - V|\sigma$ or $V_h|\sigma$ for $h \neq i$ and $h \notin S$, then $b := b + 1$ (another ballot is cast) and if the label of the $jth$ nearest neighbor is not equal to $y$, then $d := d + 1$ (another disagreeing vote). Stop when $f = k$, and return the sample value:

$$rI(\forall h < i : v_h > 0) 2^{i-1} (-1)^{|S|} I\left(d > \frac{k}{2}\right). \tag{110}$$

This method may require up to $O(rm + k)$ computation per sample, because it is possible (though extremely unlikely) for an example to have all validation examples in $V|\sigma$ as nearer neighbors than the $kth$ nearest neighbor from $F - V|\sigma$. To reduce worst-case computation, select a value $w > k$, stop computation for a sample if $w$ neighbors are assigned to subsets $V_1, \ldots, V_r, F - V|\sigma$ before the $kth$ neighbor is assigned to $F - V|\sigma$ (that is, if $f < k$ and $j = w$), and return zero as the value for the sample. Then only the $w$ nearest neighbors to each example need to be found, and the remaining computation is $O(w)$ per sample. Call this the modified sampling procedure. We can use it as the basis for a bound that is feasible to compute:

**Theorem 13** *Let $\hat{p}_s$ be the empirical mean of $s$ i.i.d. samples of the modified sampling procedure. Let $\hat{s}$ represent the modified sampling procedure. Then*

$$\forall \delta > 0 : Pr_{F \sim D^n, \hat{s}} \left\{ |p* - \hat{p}_s| \geq \epsilon_v + \epsilon_r + \epsilon_c + \epsilon_s \right\} \leq \delta, \tag{111}$$

*where*

$$\epsilon_v = \sum_{i=1}^{r} 2^{i-1} Pr_{\sigma \in P} \{b_{i-1}|\sigma\} \frac{1}{\sqrt{2m}} \left( 1.06 \sqrt{\ln \frac{20}{9\delta}} + 3.22 \right), \tag{112}$$

$$\epsilon_r = 2^{r-1} Pr_{\sigma \in P} \{b_r|\sigma\}, \tag{113}$$

$$\epsilon_c = r 2^{r-1} \sum_{i=0}^{k-1} \binom{w}{i} \frac{\prod_{j=0}^{i-1}(n - rm - j) \prod_{j=0}^{w-i-1}(rm - 1 - j)}{\prod_{j=0}^{w-1}(n - 1 - j)}, \tag{114}$$

*and*

$$\epsilon_s = \sqrt{\frac{2v \ln \frac{20}{\delta}}{s}} + \frac{r 2^r \ln \frac{20}{\delta}}{3s}, \tag{115}$$

*where*

$$v = \frac{1}{r} \sum_{i=1}^{r} Pr_{\sigma \in P} \{b_{i-1}|\sigma\} r^2 2^{2(i-1)}. \tag{116}$$

***Proof*** Let $p_c$ be $p_I$, but with terms set to zero if $\sigma$ places fewer than $k$ of the $w$ nearest neighbors to $x$ from $F$ into $F - V|\sigma$. That is, $p_c$ is $p_I$, but with terms set to zero if they are set to zero in modified sampling. Then

$$p^* = p_c + (p_I - p_c) + p_{II}. \tag{117}$$

Let $\hat{p}_c$ be $\hat{p}_I$, but with terms set to zero if they are set to zero in modified sampling. Then $\hat{p}_c$ is an unbiased empirical-mean estimate of $p_c$, and

$$\hat{p}_s = \hat{p}_c + (\hat{p}_s - \hat{p}_c). \tag{118}$$

So

$$|p* - \hat{p}_s| = |[p_c + (p_I - p_c) + p_{II}] - [\hat{p}_c + (\hat{p}_s - \hat{p}_c)]| \tag{119}$$

$$\leq |p_c - \hat{p}_c| + |p_{II}| + |p_I - p_c| + |\hat{p}_s - \hat{p}_c|. \tag{120}$$

To prove the theorem, we will show results for $\epsilon_v$, $\epsilon_r$, $\epsilon_c$, and $\epsilon_s$:

$$Pr_{F\sim D^n}\left\{|p_c - \hat{p}_c| \geq \epsilon_v\right\} \leq \frac{9}{10}\delta, \tag{121}$$

$$|p_{II}| \leq \epsilon_r, \tag{122}$$

$$|p_I - p_c| \leq \epsilon_c, \tag{123}$$

and

$$Pr_{\hat{s}}\left\{|\hat{p}_s - \hat{p}_c| \geq \epsilon_s\right\} \leq \frac{1}{10}\delta. \tag{124}$$

For $\epsilon_v$, $p_c$ and $\hat{p}_c$ are $p_I$ and $\hat{p}_I$, with some $f_{i,\sigma}$ set to zero, which can only reduce the ranges of the terms in $p_Q$ and $\hat{p}_Q$. So the result from Theorem 8 (Inequality 64):

$$\forall\delta > 0 : Pr_{F\sim D^n}\left\{|p_I - \hat{p}_I| \leq \epsilon_I\right\} \leq \delta \tag{125}$$

also applies with $p_c$ and $\hat{p}_c$ in place of $p_I$ and $\hat{p}_I$. We use a probability of bound failure $\frac{9}{10}\delta$ in place of $\delta$, so $\epsilon_v$ is $\epsilon_I$, with $\frac{9}{10}\delta$ in place of $\delta$. (We preserve the other $\frac{1}{10}\delta$ probability of bound failure for using $\hat{p}_s$ to estimate $\hat{p}_c$.) So

$$Pr_{F\sim D^n}\left\{|p_c - \hat{p}_c| \geq \epsilon_v\right\} \leq \frac{9}{10}\delta. \tag{126}$$

For $\epsilon_r$, $\epsilon_r = \epsilon_{II}$, so apply Inequality 74 directly:

$$|p_{II}| \leq \epsilon_r. \tag{127}$$

For $\epsilon_c$, we need the probability that fewer than $k$ of the nearest $w$ neighbors to an $(x, y) \in V|\sigma$ from $F - (x, y)$ are in $F - V|\sigma$. The probability that the $i$ nearest neighbors are in $F - V|\sigma$ is

$$\left(\frac{n - rm}{n - 1}\right)\left(\frac{n - rm - 1}{n - 2}\right)\cdots\left(\frac{n - rm - (i - 1)}{n - i}\right). \tag{128}$$

Given this, the probability that the next $w - i$ nearest neighbors are in $V - (x, y)|\sigma$ is

$$\left(\frac{rm - 1}{n - i - 1}\right)\left(\frac{rm - 2}{n - i - 2}\right)\cdots\left(\frac{rm - (w - i)}{n - w}\right). \tag{129}$$

So take the product of these two products. There are $\binom{w}{i}$ different ways to choose positions for the $i$ neighbors in $F - V|\sigma$ among the first $w$ neighbors. Each set of positions has equal

probability. So multiply by $\binom{w}{i}$. Sum over $i < k$ and multiply by the maximum term value to get $\epsilon_c$. So

$$|p_I - p_c| \leq \epsilon_c. \tag{130}$$

For $\epsilon_s$, apply a result from Maurer and Pontil (2009) (page 2, Theorem 3) derived from Bennett's Inequality (Bennett 1962), on the difference between the mean $\mu$ of a distribution and an empirical mean $\hat{\mu}$ over $s$ samples drawn i.i.d. according to the distribution:

$$\forall \delta > 0 : Pr\left\{|\mu - \hat{\mu}| \geq \epsilon\right\} \leq \delta, \tag{131}$$

where

$$\epsilon = \sqrt{\frac{2v \ln \frac{2}{\delta}}{s}} + \frac{q \ln \frac{2}{\delta}}{3s}, \tag{132}$$

$q$ is the range of the distribution, and $v$ is any upper bound on the variance of the distribution. The result is stronger than Hoeffding bounds when sample variance is small relative to the range. To get $\epsilon = \epsilon_s$, apply this inequality with $\mu = \hat{p}_c$, $\hat{\mu} = \hat{p}_s$, $\delta$ set to $\frac{1}{10}\delta$, $q = r2^r$, and $v$ set to an upper bound on the expectation of the square of sample values. To get such an upper bound, start with the sample values:

$$rI(b_{i-1})2^{i-1}(-1)^{|S|}I(\overline{g_{-(S\cup\{i\})}}|\sigma), \tag{133}$$

drop $I(\overline{g_{-(S\cup\{i\})}}|\sigma)$, and take the expectation of the square. □

Similar to the result from Theorem 8:

**Theorem 14** *For some choices of $m$, $r$, $w$, and $s$,*

$$\epsilon_v + \epsilon_r + \epsilon_c + \epsilon_s \in O\left(\sqrt{(k + \ln n)/n}\right). \tag{134}$$

***Proof*** Apply the alternative proof of Corollary 12, with $\epsilon_v$ and $\epsilon_r$ in place of $\epsilon_I$ and $\epsilon_{II}$, keeping $r = \lceil \ln \sqrt{n} \rceil$, but with 3 in place of 2 in the value for $m$ from Equality 98. Then

$$m = \left\lfloor \frac{n}{3e^2(k + r - 1)} \right\rfloor, \tag{135}$$

and the bound on $\epsilon_{II}$ in the alternative proof becomes:

$$\epsilon_v \leq \left(\frac{2e(k + r - 1)m}{n}\right)^r \leq \frac{1}{1.5^r e^r} \in \mathrm{O}\left(\frac{1}{\sqrt{n}}\right), \tag{136}$$

Changing from 2 to 3 in $m$ only affects constants in the alternative proof's result for $\epsilon_I$, so:

$$\epsilon_v \in \mathrm{O}\left(\sqrt{(k + \ln n)/n}\right). \tag{137}$$

For $\epsilon_c$, recall that it is the maximum (absolute) value of a term times the probability that a random $\sigma \in P$ places more than $w - k$ of the $w$ closest neighbors to an $(x, y) \in V|\sigma$ from $F - (x, y)$ into $V - (x, y)|\sigma$. In Lemma 9, $d_i(x|\sigma)$ is the condition that the closest $k + i - 1$ neighbors in $F$ include at least $i$ in $V|\sigma$. So apply the result from Lemma 9:

$$Pr_{\sigma \in P}\{d_i|\sigma\} \leq \left(\frac{(k + i - 1)m}{n}\right)^i e^i, \tag{138}$$

with $k+i-1=w$ and $i=w-k+1$. (Using $F-(x,y)$ in place of $F$ and $V-(x,y)|\sigma$ in place of $V|\sigma$ can only decrease the probability, so just use $m$ and $n$ as in the lemma for this bound.) Let $w=k+r-1$. Then the RHS is

$$\leq \left(\frac{(k+r-1)m}{n}\right)^r e^r. \tag{139}$$

Since each term is less than $r2^r$,

$$\epsilon_c \leq r\left(\frac{2e(k+r-1)m}{n}\right)^r. \tag{140}$$

Using Inequality 136,

$$\epsilon_c \leq \frac{r}{1.5^r e^r}. \tag{141}$$

Note that $r<1.5^r$. (To prove it, show that $r^{1/r}<1.5$ by setting the derivative of $\ln(r^{1/r})=(1/r)\ln r$ to zero. Solve: $r=e$. So $\max r^{1/r}=e^{1/e}<1.5$.) Then

$$\epsilon_c \leq \frac{1}{e^r}. \tag{142}$$

With $r=\lceil \ln\sqrt{n}\rceil$,

$$\epsilon_c < \frac{1}{\sqrt{n}} \in \mathrm{O}\left(\frac{1}{\sqrt{n}}\right). \tag{143}$$

We can make $\epsilon_s$ arbitrarily small by increasing $s$. How many samples we need to achieve a given $\epsilon_s$ depends on $v$. To bound $v$, note that $b_{i-1}$ has probability at most

$$\left(\frac{e(k+i-2)m}{n}\right)^{i-1}, \tag{144}$$

according to Lemma 9. So

$$v=\frac{1}{r}\sum_{i=1}^{r} Pr_{\sigma\in P}\{b_{i-1}|\sigma\}\, r^2 2^{2(i-1)} \tag{145}$$

$$\leq \frac{1}{r}\sum_{i=1}^{r}\left(\frac{e(k+i-2)m}{n}\right)^{i-1} r^2 2^{2(i-1)} \tag{146}$$

$$= r\sum_{i=1}^{r}\left(\frac{4e(k+i-2)m}{n}\right)^{i-1}. \tag{147}$$

So set

$$m < \frac{n}{4e(k+r-2)} \tag{148}$$

to ensure $v\leq r^2$. (The value for $m$ in Equality 135 meets this condition.) Let $s=rn$. Then

$$\epsilon_s \leq \sqrt{\frac{2r\ln\frac{20}{\delta}}{n}} + \frac{2^r \ln\frac{20}{\delta}}{3n}. \tag{149}$$

With $r=\lceil \ln\sqrt{n}\rceil$,

$$\epsilon_s \leq \sqrt{\frac{(\ln n)\left(\ln\frac{20}{\delta}\right)}{n}} + \frac{\ln\frac{20}{\delta}}{3\sqrt{n}}. \tag{150}$$

So

$$\epsilon_s \in \mathrm{O}\left(\sqrt{(\ln n)/n}\right). \tag{151}$$

□

Appendix A presents methods to compute rather than estimate $\hat{p}_I$ or $\hat{p}_c$. The method to compute $\hat{p}_c$ requires $\mathrm{O}(n \ln n)$ computation, like sampling, but it requires $\mathrm{O}((\ln n)^4)$ space, and it is more complicated than sampling.

## 5 Tests

To apply the error bound method from the previous section to some actual classifiers, we use three randomly generated in-sample data sets of different sizes. Each example input is drawn uniformly at random from $[-1, 1]^3$, the label is set to one if an even number of coordinates are negative and zero otherwise, giving each quadrant in the cube a different label than the quadrants bordering its sides. Then, to add some noise, with probability $\frac{1}{10}$ the label is changed: from one to zero or zero to one.

In this section, we compute bounds for specific values of $n$, $k$, and $\delta$ rather than prove asymptotic results. So we use a tighter version of $\epsilon_v$, instead of the more easily analyzed, but looser, form in Theorem 13. We use the dynamic programming procedure from an appendix of Bax and Kooti (2016), with their parameter $t = 3$. The resulting bounds are about 1.5 times the corresponding Hoeffding bounds and about half the bounds using the value for $\epsilon_v$ from the previous section. Call the optimized version $\epsilon'_v$.

For each bound we use $\epsilon_r$, $\epsilon_c$, and $\epsilon_s$ and the estimation procedure from Theorem 13. We set $w = 29$, $s = 10$ million, and $\delta = 0.05$. With $w = 29$, $\epsilon_c < 0.000005$ for all bounds we computed, so we do not show it in the tables. (It would be displayed as 0.00000 for all entries.)

Tables 1, 2, and 3 show error bounds for the three data sets, with $n = 20{,}000$, $n = 50{,}000$, and $n = 100{,}000$, respectively. For each data set and $k \in \{3, 5, 7, 9, 11, 13\}$, we minimize $\epsilon'_v + \epsilon_r + \epsilon_c + \epsilon_s$ over $(r, m) \in \{1, 2, \ldots, 10\} \times \{0.001n, 0.002n, \ldots, 0.099n\}$. With the minimizing $(r, m)$, we then use the data to perform the sampling procedure to compute $\hat{p}_s$.

For each $k$, each table shows the minimizing values of $r$ and $m$, the values of $\epsilon'_v$, $\epsilon_r$, $\epsilon_c$, and $\epsilon_s$, the sample-based estimate $\hat{p}_s$, and the sum $\epsilon = \epsilon'_v + \epsilon_r + \epsilon_c + \epsilon_s$. The bound is $\hat{p}_s \pm \epsilon$. For comparison, we also include an estimate of the out-of-sample error rate $p^*$, which is the

**Table 1** For $n = 20{,}000$ and $k \in \{3, 5, \ldots, 13\}$, error bound ranges $\epsilon$ for $\epsilon$-minimizing $r$ and $m$

| $k$ | $r$ | $m$ | $\epsilon'_v$ | $\epsilon_r$ | $\epsilon_s$ | $\hat{p}_s$ | $\epsilon$ | est. $p^*$ |
|---|---|---|---|---|---|---|---|---|
| 3 | 4 | 660 | 0.09164 | 0.00365 | 0.00296 | 0.1431 | 0.0982 | 0.1467 |
| 5 | 5 | 480 | 0.11295 | 0.00192 | 0.00362 | 0.1313 | 0.1185 | 0.1331 |
| 7 | 5 | 380 | 0.12943 | 0.00199 | 0.00370 | 0.1249 | 0.1351 | 0.1283 |
| 9 | 5 | 320 | 0.14378 | 0.00220 | 0.00379 | 0.1223 | 0.1498 | 0.1265 |
| 11 | 5 | 260 | 0.15750 | 0.00175 | 0.00373 | 0.1239 | 0.1630 | 0.1253 |
| 13 | 5 | 240 | 0.16891 | 0.00232 | 0.00388 | 0.1204 | 0.1751 | 0.1248 |

Bound range increases with $k$. Bound is $\hat{p}_s \pm \epsilon$. Error estimate $\hat{p}_s$ is close to (estimated) actual error rate $p^*$ (Estimated $p^*$ is based on 10 million out-of-sample examples)

**Table 2** Error bound ranges $\epsilon$ and error bounds $\hat{p}_s \pm \epsilon$ for $n = 50{,}000$

| $k$ | $r$ | $m$ | $\epsilon_v'$ | $\epsilon_r$ | $\epsilon_s$ | $\hat{p}_s$ | $\epsilon$ | est. $p^*$ |
|---|---|---|---|---|---|---|---|---|
| 3 | 5 | 1550 | 0.06042 | 0.00135 | 0.00338 | 0.1435 | 0.0651 | 0.1421 |
| 5 | 5 | 1150 | 0.07275 | 0.00155 | 0.00354 | 0.1287 | 0.0778 | 0.1270 |
| 7 | 5 | 900 | 0.08330 | 0.00152 | 0.00360 | 0.1227 | 0.0884 | 0.1219 |
| 9 | 5 | 750 | 0.09245 | 0.00161 | 0.00366 | 0.1211 | 0.0977 | 0.1199 |
| 11 | 5 | 650 | 0.10064 | 0.00175 | 0.00373 | 0.1208 | 0.1061 | 0.1190 |
| 13 | 5 | 550 | 0.10860 | 0.00153 | 0.00370 | 0.1209 | 0.1138 | 0.1185 |

Increased $n$ (vs. 20,000 in previous table) decreases error bound ranges

**Table 3** Results for $n = 100{,}000$

| $k$ | $r$ | $m$ | $\epsilon_v'$ | $\epsilon_r$ | $\epsilon_s$ | $\hat{p}_s$ | $\epsilon$ | est. $p^*$ |
|---|---|---|---|---|---|---|---|---|
| 3 | 5 | 3000 | 0.04368 | 0.00114 | 0.00333 | 0.1386 | 0.0481 | 0.1384 |
| 5 | 5 | 2200 | 0.05251 | 0.00124 | 0.00346 | 0.1228 | 0.0572 | 0.1226 |
| 7 | 5 | 1700 | 0.06010 | 0.00115 | 0.00350 | 0.1187 | 0.0648 | 0.1175 |
| 9 | 5 | 1400 | 0.06669 | 0.00115 | 0.00354 | 0.1166 | 0.0714 | 0.1157 |
| 11 | 5 | 1200 | 0.07257 | 0.00119 | 0.00358 | 0.1158 | 0.0773 | 0.1149 |
| 13 | 5 | 1100 | 0.07761 | 0.00153 | 0.00370 | 0.1153 | 0.0828 | 0.1145 |

For each $k$, error bound ranges $\epsilon$ are about half those for $n = 20{,}000$

average over 10 million out-of-sample examples drawn i.i.d. from the same distribution as the in-sample examples.

Overall, the error estimates $\hat{p}_s$ are close to the estimated out-of-sample error rates $p^*$, with mean absolute differences 0.29% for $n = 20{,}000$, 0.12% for $n = 50{,}000$, and 0.07% for $n = 100{,}000$. The error bound ranges $\epsilon$ range from about 5% to about 17.5%, growing with $k$ and shrinking as the number of in-sample examples increases. The optimal $r$ value is 4 for $k = 3$ and $n = 20{,}000$ and 5 for the other bounds. This shows that moving beyond the previous $r = 2$-style bounds (Bax 2012) can strengthen bounds, even for moderate numbers of in-sample examples.

In general, $\epsilon_v'$ is the main contributor to error bound range $\epsilon$, with $\epsilon_r$ and $\epsilon_s$ contributing less than 0.4% in every case. The small contributions from $\epsilon_r$ may seem surprising, since the choice of $m$ mediates a tradeoff between $\epsilon_v'$ and $\epsilon_r$. However, increasing $m$ decreases $\epsilon_v'$ slowly (approximately $\mathrm{O}\left(1/\sqrt{m}\right)$) but increases $\epsilon_r$ quickly (as $\mathrm{O}\left(m^r\right)$). Optimizing $m$ means balancing the derivatives with respect to $m$ of $\epsilon_v'$ and $\epsilon_r$, not their values, and this occurs at a large value of $\epsilon_v'$ relative to $\epsilon_r$.

The optimal values of $m$ are small relative to $n$. They range from about 1% of the data to about 3%. With $r = 5$, that is about 5% to 15% of the in-sample examples in $V$. The fraction shrinks as $k$ increases, because $m$ must be smaller to avoid having a large probability of all validation data sets having examples closer to a random input than the $k$ closest examples in $F - V$—to avoid a large $\epsilon_r$. The small values of $m$ still produce moderately small $\epsilon_v'$, because we are using a bound for a single estimate rather than a uniform bound over error estimates for a large class of classifiers as is the case for traditional VC-style bounds (Vapnik and Chervonenkis 1971).

## 6 Conclusion

We have shown that $k$-nearest neighbor classifiers have exponential PAC error bounds with

$$\mathrm{O}\left(\sqrt{(k+\ln n)/n}\right) \tag{152}$$

error bound ranges. The bounds are quite general. They apply to any type of inputs, because they are based on probability rather than geometry. As a result, they have no terms that increase with the number of dimensions or other properties of the input space. The bounds do not require the $k$-nn classifier's method to compute distances among examples to be symmetric or to obey the triangle inequality—it need not be a metric in the mathematical sense. It can be any function on two example inputs that returns a number.

We average bounds over all choices of validation subsets so that we can prove the resulting bound has a small range. If, instead, we use a single random choice of validation subsets, then we can also produce an exponential PAC error bound. To do this, use each validation subset $V_i$ to validate $f_i()$, and use a random subset of the remaining in-sample examples to validate the rate of all validation subsets having a neighbor closer to an input than the $kth$ nearest neighbors among the other in-sample examples. (In a transductive setting (Vapnik 1998), or if unlabeled inputs are otherwise available, use them for this validation.) This bound has $\mathrm{O}\left(\sqrt{(k+\ln n)/n}\right)$ range in expectation. We average over choices of validation subsets to guarantee that we realize the expectation.

We use bounds on $Pr_{\sigma\in P}\{b_{i-1}|\sigma\}$ to bound the range of the random variables in $\epsilon_v$ and to bound the variance in $\epsilon_s$. If the classifier is accurate (and the votes are mostly not near-ties), then $I(\overline{g_{-(S\cup\{i\})}}|\sigma)$ tends to be zero for a large portion of $(S,i,\sigma)$. So the variance among terms in $\hat{p}_I$ and among terms in $\hat{p}_s$ tends to be very small. In those cases, using empirical Bernstein bounds (Audibert 2004), such as those by Maurer and Pontil (2009) (Theorem 3, page 2), can significantly shrink $\epsilon_v$ and $\epsilon_s$, because those bounds scale with $\sqrt{\hat{v}/m}$, where $\hat{v}$ is the sample variance. To shrink the variance in those cases, tighten $a_i$ to be the RHS of Expression 23. We can use the resulting definition of $b_{i-1}$ to validate terms in $p_I$, but still need to use $b_r$ as defined in this paper to bound $p_{II}$, keeping $\epsilon_{II}$ and $\epsilon_r$ the same.

We showed how to use sampling to "estimate the estimates" of the error bounds. We also showed (in the appendix) an efficient, but more complex and space-consuming, method to compute an estimate. It may be possible to improve or simplify that procedure by gathering terms in a different way. In the future, it would be interesting to explore how close the estimate developed in this paper tends to be to actual error rate for practical problems, and whether it tends to outperform the leave-one-out estimate.

It would be interesting to extend the $k$-nearest neighbor error bounds from this paper to cover selection of a distance metric from a parameterized set of "hypothesis" metrics (Kedem et al. 2012). One approach might be to use uniform bounds of the type derived in this paper over the class of potential metrics. The bounds might depend on some notion of the complexity of the class of potential metrics.

Finally, it would be interesting to apply the speculate-correct technique from this paper to derive error bounds for classifiers other than nearest neighbors. Other local classifiers include some collective classifiers (Sen et al. 2008; Macskassy and Provost 2007), such as network classifiers based only on neighbors or neighbors of neighbors in a graph. (For some background on error bounds for network classifiers, refer to London et al. (2012), Li et al. (2012) and Bax et al. (2013)). It may also be possible to apply the speculate-correct method to other types of classifiers that are typically based on small subsets of the in-sample examples,

such as support vector machines (Vapnik 1998; Cristianini and Shawe-Taylor 2000; Joachims 2002) and set-covering machines (Marchand and Shawe-Taylor 2001).

**Acknowledgements** We thank the anonymous referees for their detailed and extremely helpful corrections on the main results and advice on testing and presentation.

## A Method to compute $\hat{p}_I$ and $\hat{p}_c$

By gathering terms rather than sampling, we can compute $\hat{p}_I$ and $\hat{p}_c$ exactly. In this appendix, we show how to compute $\hat{p}_I$ exactly and how to compute $\hat{p}_c$ in $O(n \ln n)$ time and $O((\ln n)^4)$ space, assuming $w = k+r-1$ and $k+r \in O(\ln n)$, and ignoring any time and space required to find the $k+r$ nearest neighbors to each in-sample example. The methods in this section are inspired by a similar approach for a single validation subset by Mullin and Sukthankar (2000).

Recall from Eqs. 105 and 106 that

$$\hat{p}_I = E_{(x,y)\in F}\left\{rE_{i\in R}\left\{E_{\sigma\in P((x,y),i)}\left\{f_{i,\sigma}\right\}\right\}\right\}, \tag{153}$$

and

$$f_{i,\sigma} = I(b_{i-1}|\sigma)2^{i-1}E_{S\subseteq A_{i-1}}\left\{(-1)^{|S|}I(\overline{g_{-(S\cup\{i\})}}|\sigma)\right\}. \tag{154}$$

Use the symmetry of permutations over same-size subsets $S$ to compute only for $S = \{1,\ldots,|S|\}$, and use $s$ to index values of $|S|$. Note that

$$Pr_{S\subseteq A_{i-1}}\{|S| = s\} = \binom{i-1}{s}2^{-(i-1)}. \tag{155}$$

Let $A_s = \{1,\ldots,s\}$. Let

$$p_{(x,y),i} = Pr_{\sigma\in P((x,y),i)}\left\{b_{i-1}\wedge\overline{g_{-A_s\cup\{i\}}}|\sigma\right\}. \tag{156}$$

Then

$$\hat{p}_I = E_{(x,y)\in F}\left\{\sum_{i=1}^{r}\sum_{s=0}^{i-1}\binom{i-1}{s}(-1)^s p_{(x,y),i}\right\}. \tag{157}$$

Refer to the $jth$ nearest neighbor to $(x, y)$ in $F-\{(x, y)\}$ as neighbor $j$. Let $c_{t,u,v}(\sigma)$ be the condition that a permutation $\sigma$ assigns the neighbors to $(x, y)$ to sets $F, V_1, \ldots, V_r|\sigma$ such that there are exactly $k$ voters (in $F-(V_1\cup\ldots\cup V_s\cup V_i)|\sigma$) among neighbors 1 to $t$, neighbor $t$ is a voter, there are $k$ neighbors from $F-V|\sigma$ among neighbors 1 to $u$, neighbor $u$ is from $F-V|\sigma$, and there are $v$ voters among neighbors 1 to $u$. Let

$$p_{t,u,v} = Pr_{\sigma\in P((x,y),i)}\left\{c_{t,u,v}(\sigma)\right\}, \tag{158}$$

and

$$\hat{P} = \{\sigma\in P((x, y), i) : c_{t,u,v}(\sigma)\}. \tag{159}$$

Then

$$p_{(x,y),i} = \sum_{t=k}^{k+(s+1)m-1}\sum_{u=t}^{k+rm-1}\sum_{v=k}^{u} p_{t,u,v}p_s p_{i-1-s}p_g, \tag{160}$$

where

$$p_s = Pr_{\sigma\in\hat{P}}\{b_s|\sigma\}, \tag{161}$$

$$p_{i-1-s} = Pr_{\sigma \in \hat{P}} \{\neg a_{s+1} \wedge \ldots \wedge \neg a_{i-1} | \sigma\}, \text{ and} \tag{162}$$

$$p_g = Pr_{\sigma \in \hat{P}} \left\{\overline{g_{-A_s \cup \{i\}}} | \sigma\right\}. \tag{163}$$

To see why, compare this to Eq. 156. For each $(t, u, v)$, we multiply the probability of $c_{t,u,v}$, which is $p_{t,u,v}$, by $p_s$, $p_{i-1-s}$, and $p_g$, each conditioned on $c_{t,u,v}$. (Taking probabilities over $\hat{P}$ conditions on $c_{t,u,v}$.) Together, the conditions in $p_s$, $p_{i-1-s}$, and $p_g$ are equivalent to the condition in Eq. 156, because $b_s \wedge \neg a_{s+1} \wedge \ldots \wedge \neg a_{i-1} | \sigma$ equals $b_{i-1} | \sigma$. The limits of summation for $t$ and $u$ follow from the fact that, with $(x, y) \in V_i | \sigma$, there are $(s+1)m - 1$ remaining non-voter assignments and $rm - 1$ remaining validation subset assignments for each $\sigma$ in $P((x, y), i)$.

Probabilities $p_s$, $p_{i-1-s}$, and $p_g$, each conditioned on $c_{t,u,v}$, are independent of each other: $c_{t,u,v}$ specifies that there are $u - v$ non-voters before the $kth$ neighbor in $F - V$, so it completely determines $p_s$. Also, $c_{t,u,v}$ specifies that neighbor $t$ is the $kth$ voter, so each size $k - 1$ subset of the first $t - 1$ is equally likely to be the other voters that determine $\overline{g_{-A_s \cup \{i\}}}$ in $p_g$, no matter how the $v$ voters are allocated among $V_{s+1} | \sigma, \ldots, V_{i-1} | \sigma$ in $p_{i-1-s}$.

To compute $p_{t,u,v}$, note that with $(x, y) \in V_i | \sigma$, for $\sigma \in P((x, y), i)$, there are $n - 1$ remaining assignments, including $m - 1$ to $V_i | \sigma$, $m$ for each other validation subset, and $n - rm$ for $F - V | \sigma$. This includes $n - (s+1)m$ voters and $(s+1)m - 1$ non-voters. So

$$p_{t,u,v} = \tag{164}$$

$$\frac{\binom{n-(s+1)m}{k}\binom{(s+1)m-1}{t-k}}{\binom{n-1}{t}} \left(\frac{k}{t}\right) \frac{\binom{n-(s+1)m-k}{v-k}\binom{(s+1)m-1-(t-k)}{u-t-(v-k)}}{\binom{n-1-t}{u-t}} \frac{\binom{n-rm}{k}\binom{(r-(s+1))m}{v-k}}{\binom{n-(s+1)m}{v}} \tag{165}$$

$$\sum_{z=0}^{\min(k, v-k)} \left(\frac{\binom{v-k}{z}\binom{k}{k-z}}{\binom{v}{k}} \frac{z}{u-t}\right). \tag{166}$$

The terms are the probabilities of the following conditions, respectively, each conditioned on the previous terms' conditions:

1. There are exactly $k$ voters among the first $t$ neighbors.
2. Neighbor $t$ is one of those $k$ voters.
3. The first $u$ neighbors include exactly $v$ voters.
4. Exactly $k$ of the $v$ voters are in $F - V | \sigma$.
5. Neighbor $u$ is from $F - V | \sigma$. (The sum is over the number $z$ of neighbors $t + 1$ to $u$ in $F - V | \sigma$.) If $\frac{z}{u-t} = \frac{0}{0}$, then treat it as one.

Now consider the three probabilities $p_s$, $p_{i-1-s}$, and $p_g$. The first is the probability that the validation subsets $V_1 | \sigma, \ldots, V_s | \sigma$ are all represented among the nearer neighbors to $(x, y)$ than the $kth$ nearest neighbor from $F - V | \sigma$. Since we condition on $c_{t,u,v}$ (by taking the probability only over $\sigma \in \hat{P}$, for which $c_{t,u,v}$ holds), the condition is that among the neighbors assigned $u - v$ of the $(s+1)m - 1$ non-voter positions, each of $s$ sets of $m$ positions is represented. Use inclusion and exclusion, counting all ways to select the $u - v$ neighbors, subtracting ways to select the $u - v$ neighbors without drawing from each set $V_1 | \sigma, \ldots, V_s | \sigma$, adding those that avoid drawing from each pair of sets, and so on:

$$p_s = \sum_{j=0}^{s} (-1)^j \binom{s}{j} \binom{(s+1-j)m - 1}{u - v} \binom{(s+1)m - 1}{u - v}^{-1}. \tag{167}$$

Similarly, the condition for $p_{i-1-s}$, given $c_{t,u,v}$, is that all of $V_{s+1} | \sigma, \ldots, V_{i-1} | \sigma$ are represented among the $v - k$ voters with positions in $V_{s+1} \cup \ldots \cup V_{i-1} \cup V_{i+1} \cup \ldots \cup V_r | \sigma$.

(The other $k$ voters are in $F - V|\sigma$.) Once again, use inclusion and exclusion:

$$p_{i-1-s} = \sum_{j=0}^{i-1-s} (-1)^j \binom{i-1-s}{j} \binom{(r-s-1-j)m}{v-k} \binom{(r-s-1)m}{v-k}^{-1}. \tag{168}$$

The condition for $p_g$, given $c_{t,u,v}$, is that at least $\frac{k+1}{2}$ of the nearest $k$ voters, of which the last is neighbor $t$, have labels that disagree with $y$. Let $y_j$ be the label of neighbor $j$. Let $d_j$ count the labels among neighbors 1 to $j$ that disagree with $y$. Use $b$ to count how many neighbors with labels that disagree with $y$ are among the $k-1$ voters nearer to $(x, y)$ than neighbor $t$. Then

$$p_g = \sum_{b=\frac{k+1}{2}-I(y_t \neq y)}^{k-1} \binom{d_{t-1}}{b} \binom{t-1-d_{t-1}}{k-1-b} \binom{t-1}{k-1}^{-1}. \tag{169}$$

Substitute Eq. 160 into Eq. 157 to get an equation for $\hat{p}_I$:

$$\hat{p}_I = E_{(x,y)\in F} \left\{ \sum_{i=1}^{r} \sum_{s=0}^{i-1} \binom{i-1}{s} (-1)^s \sum_{t=k}^{k+(s+1)m-1} \sum_{u=t}^{k+rm-1} \sum_{v=k}^{u} p_{t,u,v} p_s p_{i-1-s} p_g \right\}. \tag{170}$$

For $\hat{p}_c$ with $w = k + r - 1$, reduce the upper limits of summation for $t$ and $u$ to $k + r - 1$:

$$\hat{p}_c = E_{(x,y)\in F} \left\{ \sum_{i=1}^{r} \sum_{s=0}^{i-1} \binom{i-1}{s} (-1)^s \sum_{t=k}^{k+r-1} \sum_{u=t}^{k+r-1} \sum_{v=k}^{u} p_{t,u,v} p_s p_{i-1-s} p_g \right\}. \tag{171}$$

To compute this value, notice that only $p_g$ depends on values that are specific to each example $(x, y)$—the values $d_{t-1}$ and $I(y_t \neq y)$. Since $p_g$ only depends on those values and $t$, we can rearrange the sum:

$$\hat{p}_c = E_{(x,y)\in F} \left\{ \sum_{t=k}^{k+r-1} p_g q(t) \right\}, \tag{172}$$

where

$$q(t) = \sum_{i=1}^{r} \sum_{s=0}^{i-1} \binom{i-1}{s} (-1)^s \sum_{u=t}^{k+r-1} \sum_{v=k}^{u} p_{t,u,v} p_s p_{i-1-s}. \tag{173}$$

To compute $q(t)$, first compute and store $p_{i-1-s}$ for all feasible $(i, s, v)$ and $p_s$ for all feasible $(s, u, v)$. Next, compute and store the last term of $p_{t,u,v}$ for all feasible $(v, u-t)$, then use those values to compute and store $p_{t,u,v}$ for all feasible $(s, t, u, v)$. This requires $O(r^4)$ computation and storage. Then compute $q(t)$ for each $t \in \{k, \ldots, k + r - 1\}$ by iterating through the sums and using the pre-computed values for $p_{t,u,v}$, $p_s$, and $p_{i-1-s}$. This requires $O(r^4)$ computation.

To compute $\hat{p}_c$, first compute $p_g$ for all feasible $(t, d_{t-1}, I(y_t \neq y))$. This requires $O(rk(k + r))$ computation and $O(r(k + r))$ storage. Then, for each $(x, y) \in F$, find its $k + r - 1$ nearest neighbors in $F - V$, use the neighbors' labels to compute $d_{t-1}$ and $I(y_t \neq y)$ for $t \in \{k, \ldots, k + r - 1\}$. This requires $O(k + r)$ computation. Then compute the sum over $t$ in Eq. 172, using $d_{t-1}$ and $I(y_t \neq y)$ values to select precomputed $p_g$ values and using the precomputed $q(t)$ values. This produces a sample value for $(x, y)$. Average those sample values over $(x, y) \in F$ to compute $\hat{p}_c$.

Using this method, aside from the time to find the $k + r - 1$ nearest neighbors to each in-sample example, the time complexity is $O(\max(r^4, rk(k + r), n(k + r)))$ and the storage

complexity is $O(\max(r^4, r(k+r)))$. If $k \in O(\ln n)$ and $r \in O(\ln n)$ and $n > (\ln n)^3$, then this is $O(n \ln n)$ time and $O((\ln n)^4)$ storage.

## References


Audibert, J. -Y. (2004). *PAC-Bayesian Statistical Learning Theory*. Ph.D. thesis, Laboratoire de Probabilities et Modeles Aleatoires, Universites Paris 6 and Paris 7. http://cermis.enpc.fr/~audibert/ThesePack.zip.

Bax, E. (2008). Nearly uniform validation improves compression-based error bounds. *Journal of Machine Learning Research*, *9*, 1741–1755.

Bax, E. (2012). Validation of $k$-nearest neighbor classifiers. *IEEE Transactions on Information Theory*, *58*(5), 3225–3234.

Bax, E., & Callejas, A. (2008). An error bound based on a worst likely assignment. *Journal of Machine Learning Research*, *9*, 581–613.

Bax, E., Li, J., Sonmez, A., & Cataltepe, Z. (2013). Validating collective classification using cohorts. In *NIPS workshop on frontiers of network analysis: methods, models, and applications*.

Bax, E., & Kooti, F. (2016). Ensemble validation: Selectivity has a price, but variety is free (pg. 3, Inequalities 8 and 9). *Baylearn 2016*. https://arxiv.org/pdf/1610.01234.pdf.

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, *57*(297), 33–45.

Blum, A., & Langford, J. (2003) PAC-MDL bounds. In *Proceedings of the 16th annual conference on computational learning theory (COLT)* (pp. 344–357).

Chvátal, V. (1979). The tail of the hypergeometric distribution. *Discrete Mathematics*, *25*(3), 285–287.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other Kernel-based learning methods*. Cambridge: Cambridge University Press.

Devroye, L., & Wagner, T. (1979). Distribution-free inequalities for the deleted and holdout estimates. *IEEE Transactions on Information Theory*, *25*, 202–207.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Berlin: Springer.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York: Wiley.

Floyd, S., & Warmuth, M. (1995). Sample compression, learnability, and the Vapnik–Chervonenkis dimension. *Machine Learning*, *21*(3), 1–36.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Berlin: Springer.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, *58*(301), 13–30.

Joachims, T. (2002). *Learning to classify text using support vector machines*. London: Kluwer Academic Publishers.

Kedem, D., Tyree, S., Sha, F., Lanckriet, G. R. & Weinberger, K. Q. (2012). Non-linear metric learning. In Pereira, F., Burges, C. J. C., Bottou, L., & Weinberger, K. Q. (Eds.), *Advances in neural information processing systems* (Vol. 25, pp. 2573–2581). Curran Associates, Inc. http://papers.nips.cc/paper/4840-non-linear-metric-learning.pdf.

Langford, J. (2005). Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, *6*, 273–306.

Li, J., Sonmez, A., Cataltepe, Z., & Bax, E. (2012). Validation of network classifiers. *Structural, Syntactic, and Statistical Pattern Recognition Lecture Notes in Computer Science*, *7626*, 448–457.

Littlestone, N., & Warmuth, M. (1986). *Relating data compression and learnability*. Unpublished manuscript, University of California, Santa Cruz.

London, B., Huang, B., & Getoor, L. (2012). Improved generalization bounds for large-scale structured prediction. In *NIPS workshop on algorithmic and statistical approaches for large social networks*.

Macskassy, S. A., & Provost, F. (2007). Classification in networked data: A toolkit and a univariate case study. *Journal of Machine Learning Research*, *8*, 935–983.

Marchand, M., & Shawe-Taylor, J. (2001). Learning with the set covering machine. In *Proceedings of the eighteenth international conference on machine learning (ICML 2001)* (pp. 345–352).

Maurer, A., & Pontil, M. (2009). Empirical Bernstein bounds and sample-variance penalization. In *22nd annual conference on learning theory (COLT)*. http://www0.cs.ucl.ac.uk/staff/M.Pontil/reading/svp-final.pdf.

Mullin, M., & Sukthankar, R. (2000). Complete cross-validation for nearest neighbor classifiers. In *Proceedings of the seventeenth international conference on machine learning* (pp. 639–646).

Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., & Eliassi-Rad, T. (2008). Collective classification in network data. *AI Magazine*, *29*(3), 93–106.

Skala, M. (2013). Hypergeometric tail inequalities: Ending the insanity. arXiv arXiv:1311.5939v1. https://arxiv.org/abs/1311.5939v1.
Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27*(11), 1134–1142. https://doi.org/10.1145/1968.1972. (ISSN: 0001-0782).
Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.
Vapnik, V., & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, *16*, 264–280.