A Two-Stage Active Learning Algorithm for Nearest Neighbors

Nick Rittler nrittler@ucsd.edu
Department of Computer Science and Engineering
University of California - San Diego
La Jolla, CA 92093, USA

Kamalika Chaudhuri kamalika@cs.ucsd.edu
Department of Computer Science and Engineering
University of California - San Diego
La Jolla, CA 92093, USA

Abstract

We introduce a simple and intuitive two-stage active learning algorithm for the training of k-nearest neighbors classifiers. Under a Hölder-like smoothness condition on the conditional probability function $\mathbb{P}(Y = y|X = x)$, we provide consistency guarantees for a modified k-nearest neighbors classifier trained on samples acquired via our scheme, and show that under a margin assumption, our actively trained classifiers enjoy tighter finite sample guarantees than passively trained k-nearest neighbor classifiers.

1. Introduction

Active learning is perhaps the most studied theoretical framework allowing for the acquisition of specific types of labeled data. In active learning, the learner has access to unlabeled training data from the data distribution, and can ask an annotator to selectively label a subset of this data. A significant body of work has demonstrated the power of active learning to lower label complexities versus passive approaches, wherein every point from the underlying distribution is labeled Balcan et al. (2006); Hanneke (2007); Beygelzimer et al. (2008); Zhang and Chaudhuri (2014).

Though important results have given insight into the potential of active learning to train nonparametric classifiers Castro and Nowak (2008); Minsker (2012), there is no complete theory of active learning in these settings. For k-nearest neighbors (k-NN) in particular, the most notable contributions are Dasgupta (2012), which gives consistency guarantees for k-NN under a wide class of active strategies, but does not concern itself with sample complexities, and Kontorovich et al. (2018), which uses a compression approach to give guarantees for a 1-nearest neighbor classifier trained on a subset of an actively acquired training sample; while this work is impressive, it does not come with consistency guarantees or matching label complexity lower bounds, leaving the problem open.

In this work, we introduce a relatively straightforwards two-stage active learning algorithm for nearest neighbors, and investigate one primary setting in which it outperforms passive counterparts: when noise falls off fairly quickly around the decision boundary, similar to the noise condition of Tsybakov Audibert and Tsybakov (2005). The simple construction of our algorithm makes for a natural consistency analysis, and has connections to disagreement-based active learning techniques from the parametric setting.

2. Preliminaries

2.1 Setting

We study a binary classification setting where instances come from a metric space $(\mathcal{X}, \rho)$, and the label space is $\mathcal{Y} = \{0, 1\}$. We assume that we have access to labeled data from some data generating measure $\mathbb{P}$ on $\mathcal{X} \times \mathcal{Y}$. $\mathbb{P}$ is the product of μ, the marginal of $\mathbb{P}$ over the instance space, and $\eta : \mathcal{X} \to [0, 1]$, which denotes the conditional probability function $\mathbb{P}(Y = 1|X = x)$.

We assume that after n i.i.d. samples from $\mathbb{P}$, which are labeled by definition, we enter a second phase of the sampling. In this phase, we may specify any $R \subseteq \mathcal{X}$ with $\mu(R) > 0$, and subsequently rejection sample from R. The main sampling algorithm, formalized in Algorithm 1, makes use of this power to provide a more specialized training set for a modified NN classifer, introduced in the following.

2.2 Classifier Considered

We use a modified k-NN classifier to facilitate analysis. For a given sample of size $n + m$, and regions $R \subseteq R^+ \subseteq \mathcal{X}$ with $\mu(R^+) > 0$, we consider the modified k-NN classifier

$$g^R_{n+m,k}(x) := \begin{cases} \mathbb{1}[\frac{1}{k}\sum_{i=1}^k Y^{(i)}_{R^+}(x) \geq \frac{1}{2}], & \text{if } x \in R \\ \mathbb{1}[\frac{1}{k}\sum_{i=1}^k Y^{(i)}_{\mathbb{P}}(x) \geq \frac{1}{2}], & \text{otherwise,} \end{cases}$$

where $Y^{(i)}_{R^+}$ denotes the i^{th} nearest neighbor to x out of all of the samples drawn from the acceptance region R^+, and $Y^{(i)}_{\mathbb{P}}(x)$ the i^{th} nearest neighbor of samples drawn from $\mathbb{P}$. As a piece of related notation, we denote via $B^{(k+1)}(x)$ the closed ball centered at x with radius defined by the distance of x to it's $k+1^{st}$-NN under ρ, and $\hat{\eta}(B^{(k+1)}(x)) := \frac{1}{k}\sum_{i=1}^k Y^{(i)}_{\mathbb{P}}(x)$.

2.3 Guarantees for k-NN in the Passive Setting

The main analytic tools we use in this work are based around those introduced in Chaudhuri and Dasgupta (2014). The main idea there is that the k-NN classifier may disagree with the Bayes optimal classifier $g^*(x) := \mathbb{1}[\eta(x) \geq \frac{1}{2}]$ in a region of space where the distribution is complex enough that n samples are not sufficient. One intermediate definition is needed to introduce this region, which is referred to as the "effective decision boundary."

Definition 1 *For $p \in (0, 1]$ the "probability radius" of a point x is the real number*

$$r(x; p) := \inf\{r|\mu(B(x, r)) \geq p\}.$$

The "effective decision boundary" can then be stated as in terms of this quantity.

Definition 2 *For any $p \in (0, 1]$ and $\Delta \in (0, \frac{1}{2}]$, the "effective decision boundary" is the set*

$$\partial_{p,\Delta} := \left\{x \mid \exists r \leq r(x; p) \ s.t. \ \left|\eta(B(x, r)) - \frac{1}{2}\right| < \Delta\right\},$$

where $\eta(S) := \frac{1}{\mu(S)} \int_S \eta d\mu$, the probability of getting a 1 label conditional on an instance being in S. The utility of this definition is illustrated by the following.

Theorem 3 (Theorem 1 of Chaudhuri and Dasgupta (2014)) *Fix $\delta \in (0, 1)$, and integers $k < n + m$. If we draw $n + m$ i.i.d. samples from $\mathbb{P}$, then with probability $1 - \delta$ over the sampling,*

$$\Pr_{X\sim\mu}(g^*(X) \neq g_{n+m,k}(X)) \leq \mu(\partial_{p,\Delta}) + \delta,$$

when for a constant c_δ we have

$$p = c_\delta \cdot \frac{k}{n+m}$$
$$\Delta = \min\left(\frac{1}{2}, \sqrt{\frac{1}{k} \cdot \log(2/\delta)}\right).$$

This work largely focuses on the role of the parameter p in determining guarantees, which intuitively corresponds to investigating the "locality" of the voting procedure deciding the prediction of the classifier. As such, we will fix Δ as in the above theorem throughout.

3. Active Sampling Scheme

Algorithm 1 introduces the active sampling scheme. It is a two-phase sampling algorithm, wherein the first phase is used to determine a region of instance space to target in the second phase, and the second phase samples a mixture between rejection samples from this target region and the underlying measure. The acceptance region used in the second phase, is approximately speaking, the part of space where voting for the k-NN classifier is contentious in the following sense.

Definition 4 *Given a sample $\mathcal{S}$ of size n, the "estimated hard region" is the set*

$$\hat{H}_{n,k} := \left\{x \in \mathcal{X} \mid \left|\hat{\eta}(B^{(k+1)}(x)) - \frac{1}{2}\right| < 2\Delta + \tilde{\Delta}\right\},$$

where

$$\tilde{\Delta} := c_0 \sqrt{\frac{d \log(n) + \log(1/\delta)}{k}}.$$

Here, c_0 is a universal constant and d denotes the VC-dimension of balls in $\mathcal{X}$.

This region approximates the"effective decision boundary" from the outside with high probability when η is sufficiently smooth, and as such, cautiously partitions instance space in two: one part where we can be confident that k-NN will predict correctly, and one where we can not be. This elimination of "non-contentious" parts of space is a standard technique in disagreement-based active learning strategies. The choice of $\tilde{\Delta}$ allows for uniform convergence arguments over the estimation of conditional probabilities, which is vital for such statements about set containment to hold Balsubramani et al. (2019).

In Algorithm 1, the actual region used for rejection sampling is a larger relative of this set.

Algorithm 1 Two-Round Sampling Algorithm

```
Initialize:
    $n \in \mathbb{N}$, $m \in \mathbb{N}$, $\pi \in [0,1]$, $\mathcal{S} = \emptyset$, $i = n$
$\mathcal{S} \leftarrow \{(X_1, Y_1), \dots, (X_n, Y_n)\} \sim \mathbb{P}^{\bigotimes n}$        ▷ Take $n$ samples i.i.d. $\sim \mathbb{P}$
if $\mu(\hat{H}^{+}_{n,k}) > 0$ then
    while $i < n + m$ do
        $(X_i, Y_i) \leftarrow \emptyset$
        Sample $P \sim \mathrm{Ber}(\pi)$, $X \sim \mu$        ▷ 'Sample' means sample independently.
        if $P = 1$ then
            $X_i \leftarrow X$
            Sample $Y_i \sim \mathrm{Ber}(\eta(X))$
        else
            if $X \in \hat{H}^{+}_{n,k}$ then
                $X_i \leftarrow X$
                Sample $Y_i \sim \mathrm{Ber}(\eta(X))$
            end if
        end if
        $\mathcal{S} \leftarrow \mathcal{S} \cup \{(X_i, Y_i)\}$
        $i \leftarrow |\mathcal{S}|$
    end while
else
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{(X_{n+1}, Y_{n+1}), \dots, (X_{n+m}, Y_{n+m})\} \sim \mathbb{P}^{\bigotimes m}$
end if
```

Definition 5 *Given a sample $\mathcal{S}$ of size n, the "augmented hard region" is the set*

$$\hat{H}^{+}_{n,k} := \bigcup_{x \in \hat{H}_{n,k}} B^{(k^{+})}(x),$$

where $k^{+} := \Theta\left(\sqrt{nd\log(8/\delta)}\right)$.

The Lebesgue measurability of this union of closed balls is guaranteed by Balcerzak and Kharazishvili (1999).

4. Guarantees

Guarantees for this algorithm rely on smoothness of the conditional probability function η. The notion of smoothness under which we operate is was also introduced in Chaudhuri and Dasgupta (2014), and is designed to be evocative of traditional notions of Hölder continuity.

Definition 6 *We say the conditional probability function η is "(α, L)-smooth" in $(\mathcal{X}, \rho, \mu)$ if for all $x, x^{'} \in \mathcal{X}$,*

$$|\eta(x) - \eta(x^{'})| \leq L\mu\left(B^{o}(x, \rho(x, x^{'}))\right)^{\alpha}.$$

4.1 Consistency

A major upside of the method suggested in this paper is that consistency is almost immediate. This is because much of our analysis centers around controlling the disagreement of our modified classifier with the Bayes optimal classifier, which follows in the tradition of Chaudhuri and Dasgupta (2014).

Theorem 7 *Fix $n \in \mathbb{N}$. Suppose $(\mathcal{X}, \rho, \mu)$ satisfies the Lebesgue differentiation theorem, and the following conditions govern the growth of k_{n+m} as $m \to \infty$:*

$$k_{n+m}/\log(n+m) \to \infty, \quad k_{n+m}/(n+m) \to 0.$$

Suppose samples are drawn according to Algorithm 1. Then for any first round sample $\mathcal{S}_n$ such that the boundary of the acceptance region $\partial \hat{H}^+_{n,k}$ has $\mu(\partial \hat{H}^+_{n,k}) = 0$, it holds that

$$R(g^{\hat{H}_{n,k}}_{n+m,k}) \overset{a.s.}{\to} R^*$$

over the choice of the second m samples, where R^* is the risk of g^*.

Here, the boundary is the boundary of the set $\hat{H}^+_{n,k}$ in the real-analytic sense, i.e. the set difference between the closure and interior of the set as given by ρ.

4.2 Speedup under Margin

The gains of Algorithm 1 over passive sampling are most obvious under the β-margin condition introduced in Chaudhuri and Dasgupta (2014). Evocative of Tsybakov noise conditions, this condition describes an η that falls off from $\frac{1}{2}$ near the decision boundary.

Definition 8 *We say $\mathbb{P}$ satisfies the "β-margin condition" if there is some $C > 0$ for which*

$$\mu\left(\left\{ x : \left| \eta(x) - \frac{1}{2} \right| \le t \right\} \right) \le C t^{\beta}.$$

The following result is a corollary (to a finite sample guarantee not shown in this document) shows that under smoothness and the β-margin condition, our actively trained classifier has tighter high probability guarantees than the passive high probability guarantees of Chaudhuri and Dasgupta (2014).

Corollary 1 *Fix $\delta \in (0,1)$, and $m = \Theta(n)$. Suppose that η is (α, L)-smooth such that*

$$\Delta \ge 3L\left(\frac{k}{n} + \sqrt{\frac{16d\log(8/\delta)}{n}} \right)^{\alpha},$$

and set

$$k \propto \log(1/\delta)^{\frac{1}{1+2\alpha(1-\beta/2)}} \log(n)^{\frac{-\alpha\beta}{1+2\alpha(1-\beta/2)}} n^{\frac{2\alpha}{1+2\alpha(1-\beta/2)}}.$$

If in addition $\mathbb{P}$ satisfies the β-margin condition with $\beta < 2$, then for all $\zeta > 0$, with probability $\ge 1 - \delta$ over the draw of $n + m$ training samples from Algorithm 1,

$$\Pr_{X\sim\mu}\left(g^*(X) \ne g_{n+m,k}(X; \hat{H}_{n,k}) \right) \le O\left(\left(\frac{\log(1/\delta)^{1-\frac{\beta}{2}}}{n} \right)^{\frac{\alpha\beta}{1+2\alpha(1-\beta/2)}} \cdot n^{\zeta} \right) + \delta.$$

The corresponding passive guarantee is Chaudhuri and Dasgupta (2014) Theorem 7a, which states that for $k \propto \log(1/\delta)^{\frac{1}{1+2\alpha}} n^{\frac{2\alpha}{2\alpha+1}}$*, with probability* $\geq 1-\delta$ *over the draw of* $n+m$ *samples from* $\mathbb{P}$*,*

$$\Pr_{X\sim\mu}\left(g^*(X) \neq g_{n+m,k}(X)\right) \leq O\left(\left(\frac{\log(1/\delta)}{n}\right)^{\frac{\alpha\beta}{2\alpha+1}}\right) + \delta.$$

The improvement of our active strategy primarily comes from the shrunken denominator in the exponent of n^{-1}, but shows up in the $\log(1/\delta)$ term as well. We note that the condition that

$$\Delta \geq 3L\left(\frac{k}{n} + \sqrt{\frac{16d\log(8/\delta)}{n}}\right)^{\alpha}$$

is not stringent, as Δ falls off with $\frac{1}{k}$. In general, it is important to choose $k \in \Omega(d\log(n))$ such that $\tilde{\Delta} \downarrow 0$. The large size of k is a relic of uniform convergence arguments, and it is not yet clear to what extent it is necessary.

5. Future Work

Having explored conditions under which our active learning scheme outperforms passive schemes, natural followups include a thorough investigation of how close our guarantees come to lower bounds for active learning strategies Castro and Nowak (2008). There is also room to consider how such an algorithm can be extended to a multi-stage sampling scheme, and just how much utility further sampling stages provide. A more thorough investigation of necessity of working in a large k regime would also be ideal.

References

Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers under the margin condition. *arXiv preprint math/0507180*, 2005.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, page 65–72, 2006.

M Balcerzak and A Kharazishvili. On uncountable unions and intersections of measurable sets. 1999.

Akshay Balsubramani, Sanjoy Dasgupta, Shay Moran, et al. An adaptive nearest neighbor rule for classification. *Advances in Neural Information Processing Systems*, 32, 2019.

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. *CoRR*, abs/0812.4952, 2008.

Rui M Castro and Robert D Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. *CoRR*, 2014.

Sanjoy Dasgupta. Consistency of nearest neighbor classification under selective sampling. In *COLT*, 2012.

Steve Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning*, page 353–360, 2007.

Aryeh Kontorovich, Sivan Sabato, and Ruth Urner. Active nearest-neighbor learning in metric spaces, 2018.

Stanislav Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13(1), 2012.

Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. *CoRR*, abs/1407.2657, 2014.