The Sample Complexity of Self-Verifying Bayesian Active Learning

Liu Yang Steve Hanneke Jaime Carbonell

June 2010
CMU-ML-10-105

MACHINE LEARNING
DEPARTMENT

Carnegie Mellon

The Sample Complexity of Self-Verifying Bayesian Active Learning

Liu Yang **Steve Hanneke** **Jaime Carbonell**

June 2010
CMU-ML-10-105

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

We prove that access to the prior distribution over target functions can improve the sample complexity of self-terminating active learning algorithms, so that it is always better than the known results for prior-dependent passive learning. In particular, this is in stark contrast to the analysis of prior-independent algorithms, where there are known simple learning problems for which no self-terminating algorithm can provide this guarantee for all priors.

This research is supported by grants from the National Science Foundation.

Keywords: Bayesian Active Learning, Learning Theory, Sample Complexity

1 Introduction and Background

Active learning is a powerful form of supervised machine learning characterized by interaction between the learning algorithm and supervisor during the learning process. In this work, we consider a variant known as *pool-based* active learning, in which a learning algorithm is given access to a (typically very large) collection of unlabeled examples, and is able to select any of those examples, request the supervisor to label it (in agreement with the target concept), then after receiving the label, selects another example from the pool, etc. This sequential label-requesting process continues until some halting criterion is reached, at which point the algorithm outputs a function, and the objective is for this function to closely approximate the (unknown) target concept in the future. The primary motivation behind pool-based active learning is that, often, unlabeled examples are inexpensive and available in abundance, while annotating those examples can be costly or time-consuming; as such, we often wish to select only the informative examples to be labeled, thus reducing information-redundancy to some extent, compared to the baseline of selecting the examples to be labeled uniformly at random from the pool (passive learning).

There has recently been an explosion of fascinating theoretical results on the advantages of this type of active learning, compared to passive learning, in terms of the number of labels required to obtain a prescribed accuracy (called the *sample complexity*): e.g., [Freund et al., 1997, Dasgupta, 2004, Dasgupta, 2005, Hanneke, 2007b, Balcan et al., 2010, Balcan et al., 2006, Kääriäinen, 2006, Hanneke, 2007a,Dasgupta et al., 2008,Hanneke, 2010,Hanneke, 2009]. In particular, [Balcan et al., 2010] show that in noise-free binary classifier learning, for any passive learning algorithm for a concept space of finite VC dimension, there exists an active learning algorithm with asymptotically much smaller sample complexity for any nontrivial target concept. In later work, [Hanneke, 2009] strengthens this result by removing a certain strong dependence on the distribution of the data in the learning algorithm. Thus, it appears there are profound advantages to active learning compared to passive learning.

However, another point noted by [Balcan et al., 2010] is that active learning often suffers from a lack of internal verification of these improvements. That is, although the number of labels required to achieve good accuracy is significantly smaller than passive learning, it is often the case that the number of labels required to *verify* that the accuracy is good is not significantly improved. In particular, this phenomenon can dramatically increase the sample complexity of active learning algorithms that adaptively determine how many labels to request before terminating by maintaining an internal estimator of their own accuracy. In short, if we require the algorithm both to *learn* an accurate function and to *know* that its function is accurate, then the number of labels required by active learning is often not significantly smaller than the number required by passive learning.

We should note, however, that all of the above considerations were proven for a learning scenario in which the target concept is considered a constant, and no information about the process that generates this concept is known a priori. Alternatively, we can consider a modification of this problem, so that the target concept can be thought of as a random variable, a sample from a known distribution (called a *prior*) over the space of possible concepts. Such a setting has been studied in detail in the context of passive learning for noise-free binary classification. In particular, [Haussler et al., 1992] found that for any concept space of finite VC dimension d, for any prior and distribution over data points, $O(d/\epsilon)$ random labeled examples are sufficient for the expected error rate of the Bayes classifier produced under the posterior distribution to be at most ϵ. Furthermore, it is known that $\Omega(1/\epsilon)$ is sometimes a lower bound on the number of random labeled examples required to achieve expected error at most ϵ, by any passive learning algorithm.

In the context of active learning (again, with access to the prior), [Freund et al., 1997] analyze the *Query By Committee* algorithm, and find that if a certain information gain quantity is lower-bounded by a constant g, then the algorithm requires only $O((d/g)\log(1/\epsilon))$ labels to achieve expected error rate at most ϵ. In particular, they show that this is satisfied for homogeneous linear separators under a near-uniform prior, and a near-uniform data distribution over the unit sphere. This represents a marked improvement over the results of [Haussler et al., 1992] for passive learning. However, the condition that the information gain be lower-bounded by a constant is quite restrictive, and many interesting learning problems are precluded by this requirement.

In the present paper, we take a more general approach to the question of active learning with access to the prior. We are interested in the broad question of whether access to the prior bridges the gap between the sample complexity of *learning* and the sample complexity of learning *with verification*. Specifically, we ask the following question.

Can a prior-dependent self-terminating active learning algorithm for a concept class of finite VC dimension always achieve expected error rate at most ϵ using $o(1/\epsilon)$ label requests?

First, in Section 3, we go through a concrete example, namely interval classifiers under a uniform data density but arbitrary prior, to illustrate the general idea, and convey some of the intuition as to why one might expect a positive answer to this question. In Section 4, we present a general proof that the answer is *always* "yes." As the known results for passive learning with access to the prior are typically $\propto 1/\epsilon$ [Haussler et al., 1992], this represents an improvement over passive learning. The proof is simple and accessible, yet represents an important step in understanding the problem of self-termination in active learning algorithms, and the general issue of the complexity of verification. Also, as this is a result that does *not* generally hold for prior-independent algorithms (even for their "average-case" behavior induced by the prior) for certain concept spaces, this also represents a significant step toward understanding the inherent value of having access to the prior.

2 Definitions and Preliminaries

First, we introduce some notation and formal definitions. We denote by $\mathcal{X}$ the *instance space*, representing the range of the unlabeled data points, and we suppose a distribution $\mathcal{D}$ on $\mathcal{X}$, which we will refer to as the *data distribution*. We also suppose the existence of a sequence $X_1, X_2, \ldots$ of i.i.d. random variables, each with distribution $\mathcal{D}$, refered to as the unlabeled data sequence. For simplicity, we will suppose this sequence is essentially inexhaustible, corresponding to the practical fact that unlabeled data are typically available in abundance as they are often relatively inexpensive to obtain. Additionally, there is a set $\mathbb{C}$ of measurable classifiers $h : \mathcal{X} \to \{-1, +1\}$, refered to as the *concept space*. We denote by d the VC dimension of $\mathbb{C}$, and in our present context we will restrict ourselves to spaces $\mathbb{C}$ with $d < \infty$, refered to as a *VC class*. We also have a probability distribution π, called the *prior*, over $\mathbb{C}$, and a random variable $h^* \sim \pi$, called the *target function*; we suppose h^* is independent from the data sequence $X_1, X_2, \ldots$. For any measurable $h : \mathcal{X} \to \{-1, +1\}$, define the *error rate* $er(h) = \mathcal{D}(\{x : h(x) \neq h^*(x)\})$. So far, this setup is essentially identical to that of [Haussler et al., 1992, Freund et al., 1997].

The protocol in active learning is the following. An active learning algorithm DHM is given as input the prior π, the data distribution $\mathcal{D}$, and a value $\epsilon \in (0, 1]$. It also (implicitly) depends on the data sequence $X_1, X_2, \ldots$, and has an indirect dependence on the target function h^* via the following type of interaction. The algorithm may inspect the values X_i for any initial segment of the data sequence, select an index $i \in \mathbb{N}$ to "request" the label of; after selecting such an index, the algorithm recieves the value $h^*(X_i)$. The algorithm may then select another index, request the label, receive the value of h^* on that point, etc. This happens for a number of rounds, $N(DHM, h^*, \epsilon, \mathcal{D}, \pi)$, before eventually the algorithm halts and returns a classifier $\hat{h}$. An algorithm is said to be *correct* if $\mathbb{E}[er\left(\hat{h}\right)] \leq \epsilon$. Define the *expected sample complexity* of DHM for $(\epsilon, \mathcal{D}, \pi)$ to be $SC(\epsilon, \mathcal{D}, \pi) = \mathbb{E}[N(DHM, h^*, \epsilon, \mathcal{D}, \pi)]$.

We will be interested in proving that certain algorithms achieve a sample complexity

$$SC(\epsilon, \mathcal{D}, \pi) = o(1/\epsilon)$$

For some $(\mathbb{C}, \mathcal{D})$, it is known that there are π-indepenent algorithms (meaning the algorithm's behavior is independent of the π argument) DHM such that we always have

$$\mathbb{E}[N(DHM, h^*, \epsilon, \mathcal{D}, \pi)|h^*] = o(1/\epsilon)$$

for instance, threshold classifiers have this property under any $\mathcal{D}$, homogeneous linear separators have this property under a uniform $\mathcal{D}$ on the unit sphere in k dimensions, and intervals with positive width on $\mathcal{X} = [0, 1]$ have this property under $\mathcal{D} = \mathrm{Uniform}([0, 1])$ (see e.g., [Dasgupta, 2005]). It is straightforward to show that any such DHM will also have $SC(\epsilon, \mathcal{D}, \pi) = o(1/\epsilon)$ for every π. In particular, the law of total expectation and the dominated convergence theorem imply

$$\lim_{\epsilon \longrightarrow 0} \epsilon SC(\epsilon, \mathcal{D}, \pi) = \lim_{\epsilon \longrightarrow 0} \epsilon \mathbb{E}[\mathbb{E}[N(DHM, h^*, \epsilon, \mathcal{D}, \pi)|h^*]] = \mathbb{E}\left[\lim_{\epsilon \longrightarrow 0} \epsilon \mathbb{E}[N(DHM, h^*, \epsilon, \mathcal{D}, \pi)|h^*]\right] = 0.$$

In these cases, we can think of SC as a kind of "average-case" analysis of these algorithms. However, there are also many $(\mathbb{C}, \mathcal{D})$ for which no such π-independent algorithm exists, achieving $o(1/\epsilon)$ sample complexity for *all* priors. For instance, this is the case for $\mathbb{C}$ as the space of interval classifiers (including the empty interval) on $\mathcal{X} = [0, 1]$ under $\mathcal{D} = \mathrm{Uniform}([0, 1])$ (this essentially follows from a proof of [Balcan et al., 2010]). Thus, any general result on $o(1/\epsilon)$ expected sample complexity for π-dependent algorithms would signify that there is a real advantage to having access to the prior.

3 An Example: Intervals

In this section, we walk through a simple and intuitive example, to illustrate how access to the prior makes a difference in the sample complexity. For simplicity, in this example we will suppose the algorithm may request the label of any point in $\mathcal{X}$, not just those in the sequence $\{X_i\}$; the same ideas can easily be adapted to the setting where queries are restricted to $\{X_i\}$. Specifically, consider $\mathcal{X} = [0, 1]$, $\mathcal{D}$ uniform on $[0, 1]$, and the concept space of *interval classifiers*, where

$\mathbb{C} = \{\mathbb{I}_{[a,b]} : 0 < a \leq b < 1\}$. For each classifier $h \in \mathbb{C}$, let $w(h) = \mathbb{P}(h(x) = +1)$ (the width of the interval h).

Consider an active learning aglorithm that makes label requests at the locations (in sequence) $1/2, 1/4, 3/4, 1/8, 3/8, 5/8, 7/8, 1/16, 3/16, \dots$ until (case 1) it encounters an example x with $h^*(x) = +1$ or until (case 2) the set of classifiers $V \subseteq \mathbb{C}$ consistent with all observed labels so far satisfies $\mathbb{E}[w(h^*)|V] \leq \epsilon$ (which ever comes first). In case 2, the algorithm simply halts and returns the constant classifier that always predicts -1: call it h_-; note that $er(h_-) = w(h^*)$. In case 1, the algorithm enters a second phase, in which it performs a binary search (repeatedly querying the midpoint between the closest two -1 and $+1$ points – taking 0 and 1 as known negative points) to the left and right of the observed positive point, halting after $\log_2(2/\epsilon)$ label requests on each side; this results in estimates of the target's endpoints up to $\pm\epsilon/4$, so that returning any classifier among the set $V \subseteq \mathbb{C}$ consistent with these labels results in error rate at most ϵ; in particular, if $\tilde{h}$ is the classifer in V returned, then $\mathbb{E}[er(\tilde{h})|V] \leq \epsilon$.

Denoting this algorithm by $DHM_{[]}$, and $\hat{h}$ the classifier it returns, we have

$$\mathbb{E}\left[er\left(\hat{h}\right)\right] = \mathbb{E}\left[\mathbb{E}\left[er\left(\hat{h}\right)\middle|V\right]\right] \leq \epsilon,$$

so that the algorithm is definitely correct.

Note that case 2 will definitely be satisfied after at most $\frac{2}{\epsilon}$ label requests, and case 1 will definitely be satisfied after at most $\frac{2}{w(h^*)}$ label requests, so that the algorithm never makes more than $\frac{2}{\max\{w(h^*),\epsilon\}}$ label requests. Abbreviating $N(h^*) = N(DHM_{[]}, h^*, \epsilon, \mathcal{D}, \pi)$, we have

$$\begin{aligned}
&\mathbb{E}[N(h^*)] \\
=&\mathbb{E}[N(h^*)|w(h^*) = 0]\mathbb{P}(w(h^*) = 0) + \mathbb{E}[N(h^*)|0 < w(h^*) \leq \sqrt{\epsilon}]\mathbb{P}(0 < w(h^*) \leq \sqrt{\epsilon}) \\
&+ \mathbb{E}[N(h^*)|w(h^*) > \sqrt{\epsilon}]\mathbb{P}(w(h^*) > \sqrt{\epsilon}) \\
\leq&\mathbb{E}[N(h^*)|w(h^*) = 0]\mathbb{P}(w(h^*) = 0) + \mathbb{E}[\mathbb{E}[N(h^*)|h^*]|0 < w(h^*) \leq \sqrt{\epsilon}]\mathbb{P}(0 < w(h^*) \leq \sqrt{\epsilon}) \\
&+ \mathbb{E}[\mathbb{E}[N(h*)|h^*]|w(h^*) > \sqrt{\epsilon}]\mathbb{P}(w(h^*) > \sqrt{\epsilon}) \\
\leq&\mathbb{E}[N(h^*)|w(h^*) = 0]\mathbb{P}(w(h^*) = 0) + \frac{2}{\epsilon}\mathbb{P}(0 < w(h^*) \leq \sqrt{\epsilon}) + \frac{2}{\sqrt{\epsilon}} + 2\log_2\frac{2}{\epsilon}. \qquad (1)
\end{aligned}$$

The third term in (1) is $O(1/\sqrt{\epsilon})$, and therefore $o(1/\epsilon)$. Since $\mathbb{P}(0 < w(h^*) \leq \sqrt{\epsilon}) \to 0$ as $\epsilon \to 0$, the second term in (1) is $o(1/\epsilon)$ as well. If $\mathbb{P}(\lambda(h^*) = 0) = 0$, this completes the proof. We focus the rest of the proof on the first term in (1), in the case that $\mathbb{P}(w(h^*) = 0) > 0$, i.e. there is nonzero probability that the target h^* labels the space almost all negative. Letting V denote the subset of $\mathbb{C}$ consistent with all requested labels, note that on the event $w(h^*) = 0$, after n label requests (for n a power of 2) we have $\max_{h \in V} w(h) \leq 1/n$. Thus, for any value $w_\epsilon \in (0, 1)$, after at most $\frac{2}{w_\epsilon}$ label requests, on the event that $w(h^*) = 0$,

$$\mathbb{E}[w(h^*)|V] \leq \frac{\mathbb{E}\left[w(h^*)\mathbb{I}[w(h^*) \leq w_\epsilon]\right]}{\pi(V)} \leq \frac{\mathbb{E}\left[w(h^*)\mathbb{I}[w(h^*) \leq w_\epsilon]\right]}{\mathbb{P}(w(h^*) = 0)}. \qquad (2)$$

Now note that, by the dominated convergence theorem,

$$\lim_{w \to 0} \mathbb{E}\left[\frac{w(h^*)\mathbb{I}[w(h^*) \leq w]}{w}\right] = \mathbb{E}\left[\lim_{w \to 0} \frac{w(h^*)\mathbb{I}[w(h^*) \leq w]}{w}\right] = 0.$$

Therefore, $\mathbb{E}\left[w(h^*)\mathbb{I}[w(h^*) \leq w]\right] = o(w)$. If we define w_ϵ as the largest value of w for which $\mathbb{E}\left[w(h^*)\mathbb{I}[w(h^*) \leq w]\right] \leq \epsilon\mathbb{P}(w(h^*) = 0)$ (or, say, half the supremum if the maximum is not achieved), then we have $w_\epsilon = \omega(\epsilon)$. Combined with (2), this implies

$$\mathbb{E}[N(h^*)|w(h^*) = 0] \leq \frac{2}{w_\epsilon} = o(1/\epsilon).$$

Thus, all of the terms in (1) are $o(1/\epsilon)$, so that in total $\mathbb{E}[N(h^*)] = o(1/\epsilon)$.

In conclusion, for this concept space $\mathbb{C}$ and data distribution $\mathcal{D}$, we have a correct active learning algorithm achieving a sample complexity $SC(\epsilon, \pi) = o(1/\epsilon)$ for all priors π on $\mathbb{C}$.

4 Main Result

In this section, we present our main result: a general result stating that $o(1/\epsilon)$ expected sample complexity is always achievable by some correct active learning algorithm. This represents an improvement over the known results for passive learning with a prior: namely, $\Theta(1/\epsilon)$. As mentioned, this type of result is often not possible for algorithms lacking access to the prior π, as there are well-known problems $(\mathbb{C}, \mathcal{D})$ for which no prior-independent correct algorithm (of the self-terminating type studied here) can achieve $o(1/\epsilon)$ sample complexity for every prior π [Balcan et al., 2010]; in particular, the intervals problem studied above is one such example.

First, we have a small lemma.

Lemma 1. *For any random variable Z and sequence of functions ϕ_n such that, for some constant $c \in (0, \infty)$, $\forall z$, $\phi_n(z) = o(1/n)$ and $\forall n \in \mathbb{N}$, $\phi_n(z) \leq c/n$, there exists a sequence $\bar{\phi}_n$ such that*

$$\bar{\phi}_n = o(1/n) \quad \textit{and} \quad \mathbb{P}\left(\phi_n(Z) > \bar{\phi}_n\right) \to 0.$$

Proof. For any constant $\delta \in (0, \infty)$, we have (by Markov's inequality and the dominated convergence theorem)

$$\lim_{n\to\infty} \mathbb{P}\left(n\phi_n(Z) > \delta\right) \leq \frac{1}{\delta} \lim_{n\to\infty} \mathbb{E}[n\phi_n(Z)] = \frac{1}{\delta}\mathbb{E}[\lim_{n\to\infty} n\phi_n(Z)] = 0.$$

Therefore (by induction), there exists a diverging sequence n_i in $\mathbb{N}$ such that

$$\sup_{n\geq n_i} \mathbb{P}\left(n\phi_n(Z) > 2^{-i}\right) \to 0$$

as $i \to \infty$. Inverting this, let $i_n = \max\{i \in \mathbb{N} : n_i \leq n\}$, and define $\bar{\phi}_n(Z) = (1/n) \cdot 2^{-i_n}$. By construction, $\mathbb{P}\left(\phi_n(Z) > \bar{\phi}_n\right) \to 0$. Furthermore, $n_i \to \infty \implies i_n \to \infty$, so that we have

$$\lim_{n\to\infty} n\bar{\phi}_n = \lim_{n\to\infty} 2^{-i_n} = 0,$$

implying $\bar{\phi}_n = o(1/n)$. □

Theorem 1. *For any VC class $\mathbb{C}$ and data distribution $\mathcal{D}$, there is a correct active learning algorithm achieving a sample complexity SC for $(\mathbb{C}, \mathcal{D})$ such that, for all priors π on $\mathbb{C}$,*

$$SC(\epsilon, \pi) = o(1/\epsilon).$$

Proof. Consider a slightly different type of active learning algorithm than that defined above: namely, an algorithm DHM_a that takes as input a *budget* $n \in \mathbb{N}$ on the number of label requests it is allowed to make, and that after making at most n label requests returns as output a classifier $\hat{h}_n$. It is known that for any $\mathbb{C}$ and $\mathcal{D}$ as above, there exists a (prior-independent) active learning algorithm DHM_a of this type, and a (target-dependent) function $R(n; h^*)$ satisfying (for some $\mathbb{C}$-dependent h^*-independent constant $c \in (0, \infty)$)

$$\forall f \in \mathbb{C}, \qquad R(n; f) \le c/n \qquad \text{and} \qquad R(n; f) = o(1/n),$$

such that we *always* have $\mathbb{E}\left[er\left(\hat{h}_n\right)\middle|h^*\right] \le R(n; h^*)$ (see [Hanneke, 2009], and related earlier work [Balcan et al., 2010]).[1] That is, equivalently, for any fixed value for the target function, the expected error rate is $o(1/n)$, where the random variable in the expectation is only the data sequence $X_1, X_2, \ldots$. Our task in this proof is to convert such an algorithm into one of the form defined in Section 1: that is, a self-terminating prior-dependent algorithm, taking ϵ as input.

For this, consider the value

$$n_\epsilon = \min\left\{n \in \mathbb{N} : \mathbb{E}\left[er\left(\hat{h}_n\right)\right] \le \epsilon\right\}.$$

This value is accessible based purely on access to π and $\mathcal{D}$. Furthermore, we clearly have (by construction) $\mathbb{E}\left[er\left(\hat{h}_{n_\epsilon}\right)\right] \le \epsilon$. Thus, denoting by DHM_a' the active learning algorithm, taking $(\mathcal{D}, \pi, \epsilon)$ as input, which runs $DHM_a(n_\epsilon)$ and then returns $\hat{h}_{n_\epsilon}$, we have that DHM_a' is a *correct* algorithm (i.e., its expected error rate is at most ϵ).

As for the expected sample complexity $SC(\epsilon, \pi, \mathcal{D})$ achieved by DHM_a', we have $SC(\epsilon, \pi, \mathcal{D}) \le n_\epsilon$, so that it remains only to bound n_ϵ. By Lemma 1, there is a π-dependent function $R(n; \pi)$ such that

$$\forall \pi, \qquad \pi\left(\{f \in \mathbb{C} : R(n; f) > R(n; \pi)\}\right) \to 0 \qquad \text{and} \qquad R(n; \pi) = o(1/n).$$

Therefore, by the law of total expectation,

$$\begin{aligned}\mathbb{E}\left[er\left(\hat{h}_n\right)\right] = \mathbb{E}\left[\mathbb{E}\left[er\left(\hat{h}_n\right)\middle|h^*\right]\right] &\le \mathbb{E}\left[R(n; h^*)\right] \\ &\le \frac{c}{n}\pi\left(\{f \in \mathbb{C} : R(n; f) > R(n; \pi)\}\right) + R(n; \pi) = o(1/n).\end{aligned}$$

In particular, this implies $n_\epsilon = o(1/\epsilon)$, as required. □

[1] Furthermore, it is not difficult to see that we can take this R to be measurable in the h^* argument.

In fact, the dependence on $\mathcal{D}$ in the algorithm described in the proof is fairly weak, and we can eliminate any direct dependence on $\mathcal{D}$ by using an algorithm DHM_a that does not depend on $\mathcal{D}$ (see [Hanneke, 2009]), and then replacing $er\left(DHM_a(n)\right)$ by a $1-\epsilon/2$ confidence upper bound based on $m_\epsilon = \Omega\left(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon}\right)$ i.i.d. unlabeled examples $X'_1, X'_2, \ldots, X'_{m_\epsilon}$ independent from the examples used by the algorithm (e.g., set aside in a pre-processing step, where the bound is calculated via Hoeffding's inequality and a union bound over the values of n that we check, of which there are at most $O(1/\epsilon)$). Then we simply increase the value of n (starting at some constant, such as 1) until

$$\frac{1}{m_\epsilon}\sum_{i=1}^{m_\epsilon}\mathbb{E}\left[\mathbb{I}\left[h^*\left(X'_i\right) \neq \left[DHM_a(n)\right]\left(X'_i\right)\right]\middle|\{X_j\}_j, \{X'_j\}_j\right] \leq \epsilon/2.$$

The expected value of the smallest value of n for which this occurs is $o(1/\epsilon)$. Note that the expectation only requires access to the prior π, not the data distribution $\mathcal{D}$; if desired for computational efficiency, this may also be estimated by a $1-\epsilon/4$ confidence upper bound based on $\Omega\left(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon}\right)$ independent samples of h^* values with distribution π, where for each sample we simulate the execution of $DHM_a(n)$ for that target function in order to obtain the returned classifier. In particular, note that no actual label requests to the oracle are required during this process of determining the appropriate label budget n_ϵ, as all executions of DHM_a are *simulated*.

References

Balcan, M.-F., Beygelzimer, A., & Langford, J. (2006). Agnostic active learning. *Proc. of the 23rd International Conference on Machine Learning (ICML).*

Balcan, M.-F., Hanneke, S., & Wortman, J. (2010). The true sample complexity of active learning. *Machine Learning Journal.*

Dasgupta, S. (2004). Analysis of a greedy active learning strategy. *Advances in Neural Information Processing Systems* (pp. 337–344). MIT Press.

Dasgupta, S. (2005). Coarse sample complexity bounds for active learning. *Proc. of Neural Information Processing Systems (NIPS).*

Dasgupta, S., Hsu, D., & Monteleoni, C. (2008). A general agnostic active learning algorithm. *Advances in Neural Information Processing Systems 20.*

Freund, Y., Seung, H. S., Shamir, E., & Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine Learning* (pp. 133–168).

Hanneke, S. (2007a). A bound on the label complexity of agnostic active learning. *Proc. of the 24th international conference on Machine learning.*

Hanneke, S. (2007b). Teaching dimension and the complexity of active learning. *Proc. of the 20th Annual Conference on Learning Theory (COLT).*

Hanneke, S. (2009). *Theoretical foundations of active learning*. Doctoral dissertation, Carnegie Mellon University.

Hanneke, S. (2010). Rates of convergence in active learning. *Annals of Statistics*.

Haussler, D., Kearns, M., & Schapire, R. (1992). Bounds on the sample complexity of bayesian learning using information theory and the vc dimension. *Machine Learning* (pp. 61–74). Morgan Kaufmann.

Kääriäinen, M. (2006). Active learning in the non-realizable case. *Proc. of the 17th International Conference on Algorithmic Learning Theory*.

MACHINE LEARNING
DEPARTMENT

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Carnegie Mellon

Carnegie Mellon University does not discriminate and Carnegie Mellon University is required not to discriminate in admission, employment, or administration of its programs or activities on the basis of race, color, national origin, sex or handicap in violation of Title VI of the Civil Rights Act of 1964, Title IX of the Educational Amendments of 1972 and Section 504 of the Rehabilitation Act of 1973 or other federal, state, or local laws or executive orders.

In addition, Carnegie Mellon University does not discriminate in admission, employment or administration of its programs on the basis of religion, creed, ancestry, belief, age, veteran status, sexual orientation or in violation of federal, state, or local laws or executive orders. However, in the judgment of the Carnegie Mellon Human Relations Commission, the Department of Defense policy of, "Don't ask, don't tell, don't pursue," excludes openly gay, lesbian and bisexual students from receiving ROTC scholarships or serving in the military. Nevertheless, all ROTC classes at Carnegie Mellon University are available to all students.

Inquiries concerning application of these statements should be directed to the Provost, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213, telephone (412) 268-6684 or the Vice President for Enrollment, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213, telephone (412) 268-2056

Obtain general information about Carnegie Mellon University by calling (412) 268-2000