A Bound on the Label Complexity of Agnostic Active Learning

Steve Hanneke[†]

March 2007
CMU-ML-07-103

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[†]Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA, shanneke@cs.cmu.edu

Abstract

We study the label complexity of pool-based active learning in the agnostic PAC model. Specifically, we derive a general upper bound on the number of label requests made by the A^2 algorithm proposed by Balcan et al. [1]. This represents the first nontrivial general-purpose upper bound on label complexity in the agnostic PAC model.

This research was sponsored through a generous grant from the Commonwealth of Pennsylvania. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the sponsoring body, or any other institution or entity.

Keywords: Agnostic Active Learning, Label Complexity, PAC

1 Introduction

In active learning, a learning algorithm is given access to a large pool of unlabeled examples, and is allowed to request the label of any particular example from that pool. The objective is to learn an accurate classifier while requesting as few labels as possible. This contrasts with passive (semi)supervised learning, where the examples to be labeled are chosen randomly. In comparison, active learning can often significantly decrease the work load of human annotators by more carefully selecting which examples from the unlabeled pool should be labeled. This is of particular interest for learning tasks where unlabeled examples are available in abundance, but labeled examples require significant effort to obtain.

In the passive learning literature, there are well-known bounds on the number of training examples necessary and sufficient to learn an accurate classifier with high probability (i.e., the sample complexity)[9, 3, 8, 2]. This quantity depends largely on the VC dimension of the concept space being learned (in a distribution-independent analysis) or the metric entropy (in a distribution-dependent analysis). However, significantly less is presently known about the analogous quantity for active learning: namely, the *label complexity*, or number of label requests that are necessary and sufficient to learn. Building a thorough understanding of label complexity, along with the quantities on which it depends, seems essential to fully exploit the potential of active learning.

In the present paper, we study the label complexity by way of bounding the number of label requests made by a recently proposed active learning algorithm, A^2 [1], which provably learns in the agnostic PAC model. The bound we find for this algorithm depends critically on a particular quantity, which we call the *disagreement coefficient*, depending on the concept space and example distribution. This quantity is often simple to calculate or bound for many concept spaces. Although we find that the bound we derive is not always tight, it represents a significant step forward, since it is the first nontrivial *general-purpose* bound on label complexity in the agnostic PAC model.

The rest of the paper is organized as follows. In Section 2, we briefly review some of the related literature, to place the present work in context. In Section 3, we continue with the introduction of definitions, notation, and some useful lemmas, along with a variety of simple examples to help build intuition. Moving on in Section 4, we state and prove the main result of this paper: an upper bound on the number of label requests made by A^2. We conclude in Section 5 with some open problems.

2 Background

The recent literature on the label complexity of active learning has been bringing us steadily closer to understanding the nature of this problem. Within that literature, there is a mix of positive and negative results, as well as a wealth of open problems.

While studying the noise-free setting, Dasgupta defines a quantity ρ called the *splitting index* [4]. ρ is dependent on the concept space, data distribution, and two (new) parameters he defines, as well as the target function itself. It essentially quantifies the amount of overlap of the "disagree sets" of (well-separated) pairs of concepts. He finds that when there is no noise, roughly $\tilde{O}(\frac{d}{\rho})$ label requests are sufficient (where d is VC dimension), and $\Omega(\frac{1}{\rho})$ are necessary for learning (for

respectively appropriate values of the other parameters). Thus, it appears that something like splitting index may be an important quantity to consider when bounding the label complexity. Unfortunately, the splitting index analysis is presently restricted to the noise-free case.

In studying the possibility of active learning in the presence of arbitrary classification noise, Balcan et al. propose the A^2 algorithm [1]. The strategy behind A^2 is to induce confidence intervals for the error rates of all concepts, and remove any concepts whose estimated error rate is larger than the smallest estimate to a statistically significant extent. This guarantees that with high probability we do not remove the best classifier in the concept space. The key observation that sometimes leads to improvements over passive learning is that, since we are only interested in *comparing* the error estimates, we do not need to request the label of any example whose label is not in dispute among the remaining classifiers. Balcan et al. analyze the number of label requests A^2 makes for some example concept spaces and distributions (notably linear separators under the uniform distribution on the unit sphere). However, other than fallback guarantees, they do not derive a general bound on the number of label requests, applicable to *any* concept space and distribution. This is the focus of the present paper.

In addition to the above results, there are a number of known lower bounds, than which there cannot be a learning algorithm guarateeing a number of label requests smaller. In particular, Kulkarni proves that, even if we allow *arbitrary* binary-valued queries and there is no noise, any algorithm that learns to accuracy $1-\epsilon$ can guarantee no better than $\Omega(\log N(2\epsilon))$ queries [7], where $N(2\epsilon)$ is the size of a minimal 2ϵ-cover (defined below). Another known lower bound is due to Kääriäinen, who proves that for most nontrivial concept spaces and distributions, if the noise rate is ν, then any algorithm that with high probability $1-\delta$ outputs a classifier with error at most $\nu+\epsilon$ can guarantee no better than $\Omega\left(\frac{\nu^2}{\epsilon^2}\log\frac{1}{\delta}\right)$ label requests [6]. In particular, these lower bounds imply that we can reasonably expect even the tightest general upper bounds on the label complexity to have some term related to $\log N(\epsilon)$ and some term related to $\frac{\nu^2}{\epsilon^2}\log\frac{1}{\delta}$.

3 Preliminaries

Let $\mathcal{X}$ be an *instance space*, comprising all possible examples we may ever encounter. $\mathbb{C}$ is a set of measurable functions $h : \mathcal{X} \to \{-1, 1\}$, known as the *concept space*.[1] $\mathcal{D}_{XY}$ is any probability distribution on $\mathcal{X} \times \{-1, 1\}$. In the active learning setting, we draw $(X, Y) \sim \mathcal{D}_{XY}$, but the Y value is hidden from the learning algorithm until requested. For convenience, we will abuse notation by saying $X \sim \mathcal{D}$, where $\mathcal{D}$ is the marginal distribution of $\mathcal{D}_{XY}$ over $\mathcal{X}$; we then say the learning algorithm (optionally) *requests* the label Y of X (which was implicitly sampled at the same time as X); we may sometimes denote this label Y by $Oracle(X)$. For any $h \in \mathbb{C}$ and distribution $\mathcal{D}'$ over $\mathcal{X} \times \{-1, 1\}$, let $er_{\mathcal{D}'}(h) = \mathcal{P}r_{(X,Y)\sim\mathcal{D}'}\{h(X) \neq Y\}$, and for $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \{-1, 1\})^m$, $er_S(h) = \frac{1}{m}\sum_{i=1}^{m} |h(x_i) - y_i|/2$. When $\mathcal{D}' = \mathcal{D}_{XY}$ (the distribution we are learning with respect to), we abbreviate this by $er(h) = er_{D_{XY}}(h)$. The *noise rate*, denoted ν, is defined as $\nu = \inf_{h\in\mathbb{C}} er(h)$. Our objective in agnostic

[1] All of the ideas described here generalize nicely to the multiclass learning task, by substituting the appropriate definitions.

active learning is to, with probability $\geq 1-\delta$, output a classifier h with $er(h) \leq \nu + \epsilon$ without making many label requests.

Let $\rho_{\mathcal{D}}(\cdot,\cdot)$ be the pseudo-metric on $\mathbb{C}$ induced by $\mathcal{D}$, such that $\forall h, h' \in \mathbb{C}, \rho_{\mathcal{D}}(h,h') = \mathcal{P}r_{X\sim\mathcal{D}}\{h(X) \neq h'(X)\}$. An ϵ-cover of $\mathbb{C}$ with respect to $\mathcal{D}$ is any set $V \subseteq \mathbb{C}$ such that $\forall h \in \mathbb{C}, \exists h' \in V : \rho_{\mathcal{D}}(h,h') \leq \epsilon$. We additionally let $N(\epsilon)$ denote the size of a minimal ϵ-cover of $\mathbb{C}$ with respect to $\mathcal{D}$. It is known that $N(\epsilon) < 2\left(\frac{2e}{\epsilon}\ln\frac{2e}{\epsilon}\right)^d$, where d is the VC dimension of $\mathbb{C}$ [5]. To focus on the learnable cases, we assume $\forall \epsilon > 0, N(\epsilon) < \infty$.

Definition 1. *The* disagreement rate $\Delta(V)$ *of a set* $V \subseteq \mathbb{C}$ *is defined as*

$$\Delta(V) = \mathcal{P}r_{X\sim\mathcal{D}}\{\exists h_1, h_2 \in V : h_1(X) \neq h_2(X)\}.$$

Definition 2. *For* $h \in \mathbb{C}$*,* $r > 0$*, let* $B(h,r) = \{h' \in \mathbb{C} : \rho_{\mathcal{D}}(h',h) \leq r\}$*. Define the* disagreement rate at radius r

$$\Delta_r = \sup_{h\in\mathbb{C}} \Delta(B(h,r)).$$

Definition 3. *The* disagreement coefficient *is the infimum value of* $\theta > 0$ *such that* $\forall r > 4(\nu+\epsilon/2)$*,*

$$\Delta_r \leq \theta r.$$

This quantity plays a critical role in the upper bounds we derive in the following section, which are increasing in this θ*.*

The canonical example of the potential improvements in label complexity of active over passive learning is the *thresholds* concept space. Specifically, consider the concept space of thresholds t_x on the interval $[0,1]$ (for $x \in [0,1]$), such that $t_x(y) = +1$ iff $y \geq x$. Furthermore, suppose $\mathcal{D}$ is uniform on $[0,1]$. In this case, it is clear that the disagreement coefficient is at most 2, since the region of disagreement of $B(t_x, r)$ is roughly $\{y \in [0,1] : |y - x| \leq r\}$. That is, since the disagreement region grows at rate 1 in two disjoint directions as r increases, the disagreement coefficient $\theta = 2$.

As a second example, let us study the disagreement coefficient for *intervals* on $[0,1]$. As before, let $\mathcal{X} = [0,1]$ and $\mathcal{D}$ be uniform, but this time $\mathbb{C}$ is the set of intervals $I_{[a,b]}$ such that for $y \in [0,1]$, $I_{[a,b]}(y) = +1$ iff $y \in [a,b]$ (for $a, b \in [0,1]$, $a \leq b$). In contrast to thresholds, the space of intervals serves as a canonical example of situations where active learning *does not help* compared to passive learning. This fact clearly shows itself in the disagreement coefficient, which is $\frac{1}{4(\nu+\epsilon/2)}$ here, since $\Delta_r = 1$ for $r = 4(\nu+\epsilon/2)$. To see this, note that the set $B(I_{[0,0]}, r)$ contains all concepts of the form $I_{[a,a]}$.

An interesting extension of the intervals example is to the space of p*-intervals*, or all intervals $I_{[a,b]}$ such that $b - a \geq p \in (0, 1/8]$. These spaces span the range of difficulty, with active learning becoming easier as p increases. This is reflected in the θ value, since here $\theta = \frac{1}{2p}$. When $r < 2p$, every interval in $B(I_{[a,b]}, r)$ has its lower and upper boundaries within r of a and b, respectively; thus, $\Delta_r \leq 4r$. However, when $r \geq 2p$, *every* interval of width p is in $B(I_{[0,p]}, r)$, so that $\Delta_r = 1$.

As an example that takes a (small) step closer to realistic learning scenarios, consider the following theorem.

Theorem 1. *If $\mathcal{X}$ is the surface of the unit sphere in $\mathbb{R}^d$, $\mathbb{C}$ is the space of homogeneous linear separators, and $\mathcal{D}$ is the uniform distribution on $\mathcal{X}$, then the disagreement coefficient θ satisfies*

$$\theta \leq \pi\sqrt{d}.$$

Proof Sketch. This result was implicitly studied by Balcan et al. [1] in their analysis of the performance of A^2 for linear separators. First we represent the concepts in $\mathbb{C}$ as weight vectors $w \in \mathbb{R}^d$ in the usual way. One can then show that $\rho_{\mathcal{D}}(w_1, w_2) = \frac{arccos(w_1 \cdot w_2)}{\pi}$. For any such w, and $r \leq 1/2$, $B(w, r) = \{w' : w \cdot w' \geq cos(r\pi)\}$. Since the hyperplane corresponding to w' is orthogonal to the vector w', the region of uncertainty swept out by this set consists of $\{x \in \mathcal{X} : |x \cdot w| \leq sin(r\pi)\}$. By geometrical considerations, this set has measure at most $\sqrt{d}sin(r\pi) \leq \pi\sqrt{d}r$. Thus, $\Delta_r \leq \pi\sqrt{d}r$. □

To prove bounds on the label complexity, we will additionally need to use some known results on finite sample rates of convergence.

Definition 4. *For $m \in \mathbb{N}$,*

$$G(m, \delta) = \sqrt{\frac{\ln \frac{2}{\delta}}{2m}}.$$

$$UB(S, h, \delta) = er_S(h) + G(|S|, \delta).$$

$$LB(S, h, \delta) = er_S(h) - G(|S|, \delta).$$

The following lemma follows immediately from Hoeffding bounds.

Lemma 1. *For $h \in \mathbb{C}$, any distribution $\mathcal{D}_i$ over $\mathcal{X} \times \{-1, 1\}$, and any $m \in \mathbb{N}$, with probability at least $1 - \delta$ over the draw of $S \sim \mathcal{D}_i^m$,*

$$|er_S(h) - er_{\mathcal{D}_i}(h)| \leq G(m, \delta).$$

In particular, this means

$$er_{\mathcal{D}_i}(h) - 2G(|S|, \delta) \leq LB(S, h, \delta) \leq er_{\mathcal{D}_i}(h) \leq UB(S, h, \delta) \leq er_{\mathcal{D}_i}(h) + 2G(|S|, \delta).$$

Furthermore, for $\gamma > 0$, if $m > \frac{1}{2\gamma^2} \ln \frac{2}{\delta}$, then $G(m, \delta) < \gamma$.

4 An Upper Bound Via the A^2 Algorithm

We use a (somewhat simplified) version of the A^2 algorithm, presented by Balcan et. al [1]. The algorithm is given in Figure 1.

The motivation behind the A^2 algorithm is to maintain a set of concepts V_i that we are confident contains any concepts with minimal error rate. If we can guarantee with statistical significance that a concept $h_1 \in V_i$ has error rate worse than another concept $h_2 \in V_i$, then we can safely remove the concept h_1 since it is suboptimal. To achieve such a statistical guarantee, the algorithm employs two-sided confidence intervals on the error rates of each classifier in the concept space; however,

Input: concept space V, accuracy parameter $\epsilon' \in (0,1)$, confidence parameter $\delta' \in (0,1)$
Output: classifier $\hat{h} \in V$
(I) $V_1 \leftarrow V$; $i \leftarrow 1$
(II) While $\Delta(V_i)\left(\min_{h \in V_i} UB(S_i, h, \delta') - \min_{h \in V_i} LB(S_i, h, \delta')\right) > \epsilon'$
1. Set $S_i = \varnothing$, $V_i' = V_i$
2. While $\Delta(V_i') \geq \frac{1}{2}\Delta(V_i)$
(a) If $\Delta(V_i')\left(\min_{h \in V_i'} UB(S_i, h, \delta') - \min_{h \in V_i'} LB(S_i, h, \delta')\right) \leq \epsilon'$
(b) Return $\hat{h} = \arg\min_{h \in V_i'} UB(S_i, h, \delta')$
(c) Else
(i) $S_i' =$ Rejection sample $2|S_i| + 1$ samples x from $\mathcal{D}$ satisfying $\exists h_1, h_2 \in V_i : h_1(x) \neq h_2(x)$
(ii) $S_i \leftarrow S_i \cup \{(x, Oracle(x)) : x \in S_i'\}$
(iii) $V_i' \leftarrow \{h \in V_i : LB(S_i, h, \delta') \leq \min_{h' \in V_i} UB(S_i, h', \delta')\}$
3. $V_{i+1} \leftarrow V_i'$; $i \leftarrow i + 1$
(III) Return $\hat{h} = \arg\min_{h \in V_i} UB(S_i, h, \delta')$

Figure 1: The A^2 algorithm.

since we are only interested in the relative *differences* between error rates, on each iteration we obtain this confidence interval for the error rate when $\mathcal{D}$ is restricted to the *region of disagreement* $\{x \in \mathcal{X} : \exists h_1, h_2 \in V_i, h_1(x) \neq h_2(x)\}$. This restriction to the region of disagreement is the primary source of any improvements A^2 achieves over passive learning. We measure the progress of the algorithm by the reduction in the measure of the region of disagreement; the key question in studying the number of label requests is bounding the number of random labeled examples from the region of disagreement that are sufficient to remove enough concepts from V_i to significantly reduce the measure of the region of disagreement.

Let V be a minimal $\frac{\epsilon}{2}$-cover of $\mathbb{C}$ with respect to $\mathcal{D}$. To obtain bounds on the label complexity, we consider running the A^2 algorithm with concept space V,[2] accuracy parameter $\epsilon' = \frac{\epsilon}{2}$, and confidence parameter $\delta' = \frac{\delta}{nN(\epsilon/2)}$, where $n = \log_2\left(\frac{512}{\epsilon^2}\ln\frac{512N(\epsilon/2)}{\epsilon^2\delta}\right)\log_2\frac{4}{\epsilon}$.

Theorem 2. *If θ is the disagreement coefficient for $\mathbb{C}$, then with probability at least $1-\delta$, given the inputs V, ϵ', and δ' described above, A^2 outputs $\hat{h} \in \mathbb{C}$ with $er(\hat{h}) \leq \nu + \epsilon$, and the total number of label requests made by A^2 is at most*

$$O\left(\theta^2\left(\frac{\nu^2}{\epsilon^2} + 1\right)\log\frac{N(\epsilon/2)\log\frac{1}{\epsilon}}{\delta}\log\frac{1}{\epsilon}\right).$$

[2] A^2 can be run with the full concept space $\mathbb{C}$ using standard uniform convergence bounds instead of those given in Lemma 1. However, we employ the $\epsilon/2$-cover to more fully exploit the distribution-dependence in this analysis. We also note that an $\epsilon/2$-cover of near optimal size can be constructed with high probability, using a polynomial number of unlabeled examples.

Proof. Let $\gamma_i = \max_{h \in V_i}(UB(S_i, h, \delta') - LB(S_i, h, \delta'))$. Since we never have $\gamma_i \leq \epsilon/4$ at step (c), Lemma 1 implies we always have $|S_i'| \leq \frac{256}{\epsilon^2} \ln \frac{2}{\delta'}$. We also never have $\Delta(V_i) \leq \epsilon/4$, so that we have at most $\log_2 \frac{4}{\epsilon}$ iterations of the outer loop. Since each iteration of the inner loop computes $er_{S_i}(h)$ for at most $N(\epsilon/2)$ concepts h, we compute at most $nN(\epsilon/2)$ such empirical errors in the entire algorithm execution. Lemma 1 and a union bound imply that, with probability $\geq 1 - \delta$, for every sample S_i formed in step (ii) of *any* iteration of the algorithm, for every $h \in V$, $|er_{S_i}(h) - er_{\mathcal{D}_i}(h)| \leq G(|S_i|, \delta')$, where $\mathcal{D}_i$ is the conditional distribution of $\mathcal{D}_{XY}$ given that $\exists h_1, h_2 \in V_i : h_1(X) \neq h_2(X)$. For the remainder of this proof, we assume that these inequalities hold for all such S_i and $h \in V$. In particular, together with the nature of the halting criterion, this implies that $er(\hat{h}) \leq \nu + \epsilon$.

Let $h^* \in V$ be such that $er(h^*) \leq \nu + \frac{\epsilon}{2}$. At step 2, suppose $\Delta(V_i) > 8\theta(\nu + \epsilon/2)$. Then let

$$V_i^{(\theta)} = \left\{ h \in V_i : \rho_{\mathcal{D}}(h, h^*) > \frac{\Delta(V_i)}{2\theta} \right\}.$$

Since for $h \in V_i$, $\rho_{\mathcal{D}}(h, h^*)/\Delta(V_i) = \rho_{\mathcal{D}_i}(h, h^*) \leq er_{\mathcal{D}_i}(h) + er_{\mathcal{D}_i}(h^*) \leq er_{\mathcal{D}_i}(h) + \frac{\nu + \epsilon/2}{\Delta(V_i)}$, we have

$$\begin{aligned}
V_i^{(\theta)} &\subseteq \left\{ h \in V_i : er_{\mathcal{D}_i}(h) > \frac{1}{2\theta} - \frac{\nu + \epsilon/2}{\Delta(V_i)} \right\} \\
&\subseteq \left\{ h \in V_i : er_{\mathcal{D}_i}(h) - \frac{1}{8\theta} > er_{\mathcal{D}_i}(h^*) + \frac{3}{8\theta} - 2\frac{\nu + \epsilon/2}{\Delta(V_i)} \right\} \\
&\subseteq \left\{ h \in V_i : er_{\mathcal{D}_i}(h) - \frac{1}{8\theta} > er_{\mathcal{D}_i}(h^*) + \frac{1}{8\theta} \right\}.
\end{aligned}$$

Let $\bar{V}_i$ denote the latter set. By Lemma 1, S_i of size $O\left(\theta^2 \log \frac{1}{\delta'}\right)$ suffices to guarantee every $h \in \bar{V}_i$ has $LB(S_i, h, \delta') > UB(S_i, h^*, \delta')$ in step (iii). Since $V_i^{(\theta)} \subseteq \bar{V}_i$ and $\Delta(V_i \setminus V_i^{(\theta)}) \leq \Delta_{\frac{\Delta(V_i)}{2\theta}} \leq \frac{1}{2}\Delta(V_i)$, we must exit the inner while loop with $|S_i| = O\left(\theta^2 \log \frac{1}{\delta'}\right)$.

On the other hand, suppose that at step 2 we have $\Delta(V_i) \leq 8\theta(\nu + \epsilon/2)$. In this case, S_i of size $O\left(\theta^2 \frac{(\nu+\epsilon)^2}{\epsilon^2} \log \frac{1}{\delta'}\right)$ suffices for every $h \in V_i$ to have $UB(S_i, h, \delta') - LB(S_i, h, \delta') < \frac{\epsilon}{2\Delta(V_i)}$, satisfying the halting conditions. Therefore, we must exit the inner while loop with $|S_i| = O\left(\theta^2 \frac{(\nu+\epsilon)^2}{\epsilon^2} \log \frac{1}{\delta'}\right)$.

These two conditions are exhaustive for any iteration of the outer while loop. Noting that there are at most $O(\log \frac{1}{\epsilon})$ iterations of the outer while loop completes the proof. □

The following lemma allows us to extend a bound for learning with $\mathcal{D}$ to a bound for any $\mathcal{D}'$ that is λ-close to $\mathcal{D}$. The proof is straightforward, and left as an exercise.

Lemma 2. *Suppose $\mathcal{D}'$ is such that, $\exists \lambda \in (0, 1)$ s.t. for all measurable sets A, $\lambda\mathcal{D}(A) \leq \mathcal{D}'(A) \leq \frac{1}{\lambda}\mathcal{D}(A)$. If Δ_r,θ,Δ_r', and θ' are the disagreement rates at radius r and disagreement coefficients for $\mathcal{D}$ and $\mathcal{D}'$ respectively, then*

$$\lambda\Delta_{\lambda r} \leq \Delta_r' \leq \frac{1}{\lambda}\Delta_{r/\lambda},$$

and thus

$$\lambda^2\theta \leq \theta' \leq \frac{1}{\lambda^2}\theta.$$

5 Open Problems

One important aspect of active learning that has not been addressed here is the value of unlabeled examples. Specifically, given an overabundance of unlabeled examples, can we use them to decrease the number of label requests required, and by how much? The splitting index bounds of Dasgupta [4] can be used to study these types of questions in the noise-free setting; however, we have yet to see a thorough exploration of the topic for agnostic learning, where the role of unlabeled examples appears fundamentally different (at least in A^2).

On the subject of bound tightness, the bound derived here for the number of label requests made by A^2 is sometimes suboptimal with respect to the label complexity of active learning. That is, there are concept spaces and distributions for which these label complexity bounds are larger than necessary (and by more than just log factors). In some cases, one can show this gap is due to a deficiency in the A^2 algorithm itself. However, in other cases, the reason for this gap remains unclear, and in particular it may be possible to derive a tighter bound for A^2 (e.g., by reducing θ^2 to θ).

Acknowledgments

I am grateful to Nina Balcan for several helpful discussions.

References

[1] M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proc. of the 23rd International Conference on Machine Learning*, 2006.

[2] G.M. Benedek and A. Itai. Learnability by fixed distributions. In *Proc. of the First Workshop on Computational Learning Theory*, pages 80–90, 1988.

[3] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.

[4] S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 18*, 2005.

[5] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

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

[7] S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis. Active learning using arbitrary binary valued queries. *Machine Learning*, 11:23–35, 1993.

[8] Sanjeev R. Kulkarni. On metric entropy, vapnik-chervonenkis dimension, and learnability for a class of distributions. Technical Report CICS-P-160, Center for Intelligent Control Systems, 1989.

[9] Vladimir Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., 1998.