Efficient Algorithms for General Active Learning

Claire Monteleoni

MIT
cmontel@csail.mit.edu

Selective sampling, a realistic active learning model, has received recent attention in the learning theory literature. While the analysis of selective sampling is still in its infancy, we focus here on one of the (seemingly) simplest problems that remain open. Given a pool of unlabeled examples, drawn i.i.d. from an arbitrary input distribution known to the learner, and oracle access to their labels, the objective is to achieve a target error-rate with minimum label-complexity, via an *efficient* algorithm. No prior distribution is assumed over the concept class, however the problem remains open even under the realizability assumption: there exists a target hypothesis in the concept class that perfectly classifies all examples, and the labeling oracle is noiseless.[1] As a precise variant of the problem, we consider the case of learning homogeneous half-spaces in the realizable setting: unlabeled examples, x_t, are drawn i.i.d. from a known distribution D over the surface of the unit ball in $\mathbb{R}^d$ and labels y_t are either -1 or $+1$. The target function is a half-space $u \cdot x \geq 0$ represented by a unit vector $u \in \mathbb{R}^d$ such that $y_t(u \cdot x_t) > 0$ for all t. We denote a hypothesis v's prediction as $v(x) = \texttt{SGN}(v \cdot x)$.

Problem: Provide an algorithm for active learning of half-spaces, such that (with high probability with respect to D and any internal randomness):

1. After L label queries, algorithm's hypothesis v obeys $P_{x \sim D}[v(x) \neq u(x)] < \epsilon$.
2. L is at most the PAC sample complexity of the supervised problem, $\tilde{O}(\frac{d}{\epsilon} \log \frac{1}{\epsilon})$, and for a general class of input distributions, L is significantly lower.[2]
3. Total running time is at most poly$(d, \frac{1}{\epsilon})$.

1 Motivation

In most machine learning applications, access to labeled data is much more limited or expensive than access to unlabeled samples from the same data-generating distribution. It is often realistic to model this scenario as active learning. Often the label-complexity, the number of labeled examples required to learn a concept via active learning, is significantly lower than the PAC sample complexity. While the query learning model has been well studied (see e.g. [1]), it is often unrealistic in practice, as it requires oracle access to the entire input space. In

[1] In the general setting, the target is the member of the concept class with minimal error-rate on the full input distribution, with respect to the (possibly noisy) oracle.

[2] Tilde notation suppresses terms in the high probability parameter, $\log d$ and $\log \log \frac{1}{\epsilon}$.

G. Lugosi and H.U. Simon (Eds.): COLT 2006, LNAI 4005, pp. 650–652, 2006.

selective sampling (originally introduced by [4]) the learner receives unlabeled data and may request certain labels to be revealed, at a constant cost per label.

2 State of the Art

Recent work has provided several negative results. Standard perceptron was shown to require $\Omega(\frac{1}{\epsilon^2})$ labels under the uniform, using any active learning rule [6]. Dasgupta [5] provided a general lower bound for learning half-spaces of $\Omega(\frac{1}{\epsilon})$ labels, when the size of the unlabeled sample is bounded. Kääriäinen provided a lower bound of $\Omega(\frac{\eta^2}{\epsilon^2})$, where η is the noise rate in the fully agnostic case [9].

Several of the positive results to date have been based on intractable algorithms. Dasgupta [5] gave a general upper bound on labels for selective sampling to learn arbitrary concepts under arbitrary input distributions, which for half-spaces under distributions λ-similar to uniform is $\tilde{O}(d \log \lambda \log^2 \frac{1}{\epsilon})$. The algorithm achieving the bound is intractable: exponential storage and computation are required, as well as access to an exponential number of functions in the concept class (not just their predictions). Similarly, recent work by Balcan, Beygelzimer and Langford [2] provides an upper bound on label-complexity of $\tilde{O}(d^2 \log \frac{1}{\epsilon})$ for learning half-spaces under the uniform, in a certain agnostic scenario, via an intractable algorithm.

Several selective sampling algorithms have been shown to work in practice, e.g. [10]. Some lack performance guarantees, or have been analyzed in the regret framework, e.g. [3]. Under a Bayesian assumption, Freund et al. [7] gave a bound on label-complexity of $\tilde{O}(d \log \frac{1}{\epsilon})$ for learning half-spaces under the uniform, using Query By Committee [13], a computationally complex algorithm that has recently been simplified to yield encouraging empirical results [8]. This is the optimal label-complexity for the problem when the input distribution is uniform, in which case the PAC sample complexity is $\tilde{\Theta}(\frac{d}{\epsilon})$ [11, 12].

There have also been some positive results for efficient algorithms, however to date the analyses have only been performed with respect to input distributions that are uniform or near-uniform. Dasgupta, Kalai and Monteleoni [6] introduced an efficient and fully online algorithm yielding the optimal label-complexity for learning half-spaces under the uniform. An algorithm due to [4], which is tractable in the realizable case, was recently shown to require at most $\tilde{O}(d^2 \log \frac{1}{\epsilon})$ labels under the uniform [2].

3 Other Open Variants

Along with the simple version stated here, the following variants remain open:

1. D is unknown to the learner.
2. Agnostic setting, under low noise rates:[3] an efficient algorithm with a non-trivial label-complexity bound under the uniform, or arbitrary distributions.

[3] The fully agnostic setting faces the lower bound of [9].

3. Online constraint: storage and time complexity (of the online update) must not scale with the number of seen labels or mistakes.
4. Analagous goal for other concept classes, or for general concepts.

References

1. D. Angluin. Queries revisited. *In Proc. 12th Int. Conference on Algorithmic Learning Theory*, LNAI,2225:12–31, 2001.
2. M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *International Conference on Machine Learning*, 2006.
3. N. Cesa-Bianchi, C. Gentile, and L. Zaniboni. Worst-case analysis of selective sampling for linear-threshold algorithms. In *Advances in Neural Information Processing Systems 17*, 2004.
4. D. A. Cohn, L. Atlas, and R. E. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.
5. S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems 18*, 2005.
6. S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Proc. 18th Annual Conference on Learning Theory*, 2005.
7. Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.
8. R. Gilad-Bachrach, A. Navot, and N. Tishby. Query by committee made real. In *Advances in Neural Information Processing Systems 18*, 2005.
9. M. Kääriäinen. On active learning in the non-realizable case. In *Foundations of Active Learning Workshop at Neural Information Processing Systems Conference*, 2005.
10. D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proc. of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, 1994.
11. P. M. Long. On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.
12. P. M. Long. An upper bound on the sample complexity of PAC learning halfspaces with respect to the uniform distribution. *Information Processing Letters*, 87(5):229–23, 2003.
13. H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proc. Fifth Annual ACM Conference on Computational Learning Theory*, 1992.