Active Learning – Modern Learning Theory

Maria-Florina Balcan and Ruth Urner
Carnegie Mellon University
Department of Machine Learning
Pittsburgh, PA, 15213, USA
ninamf@cs.cmu.edu; rurner@cs.cmu.edu

January 2015

Keywords Active learning; learning theory; sample complexity; computational complexity

Years and Authors of Summarized Original Work

2006; Balcan, Beygelzimer, Langford
2007; Balcan, Broder, Zhang
2007; Hanneke
2013; Urner, Wulff, Ben-David
2014; Awashti, Balcan, Long

Problem Definition

Most classic machine learning methods depend on the assumption that humans can annotate all the data available for training. However, many modern machine learning applications (including image and video classification, protein sequence classification, and speech processing) have massive amounts of unannotated or unlabeled data. As a consequence, there has been tremendous interest both in machine learning and its application areas in designing algorithms that most efficiently utilize the available data while minimizing the need for human intervention. An extensively used and studied technique is *active learning*, where the algorithm is presented with a large pool of unlabeled examples (such as all images available on the web) and can interactively ask for the labels of examples of its own choosing from the pool, with the goal to drastically reduce labeling effort.

Formal setup

We consider *classification problems* (such as classifying images by who is in them or classifying emails as spam or not), where the goal is to predict a label y based on its corresponding input vector x. In the standard machine learning formulation, we assume that the data points (x, y) are drawn from an unknown underlying distribution D_{XY} over $X \times Y$; X is called the feature (instance) space and $Y = \{0, 1\}$ is the label space. The goal is to output a hypothesis function h of small error (or small 0/1 loss), where $\mathrm{err}(h) = \mathbb{P}_{(x,y)\sim D_{XY}}[h(x) \neq y]$. In the passive learning setting, the learning algorithm is given a set of labeled examples $(x_1, y_1), \ldots, (x_m, y_m)$ drawn i.i.d. from D_{XY} and the goal is to output a hypothesis of small error by using only a polynomial number of labeled examples. In the *realizable* case [10](PAC learning), we assume that the true label of any example is determined by a deterministic function of the features (the so-called target function) that belongs to a known concept class C (*e.g.*, the class of linear separators, decision trees, etc). In the *agnostic* case [10, 13], we do not make the assumption that there is a perfect classifier in C, but instead we aim to compete with the best function in C (i.e., we aim to identify a classifier whose error is not much worse than $opt(C)$, the error of the best classifier in C). Both in the realizable and agnostic settings, there is a well-developed theory of sample complexity [13], quantifying in terms of the so-called *VC-dimension* (a measure of complexity of a concept class) how many training examples we need in order to be confident that a rule that does well on training data is a good rule for future data as well.

In the active learning setting, a set of labeled examples $(x_1, y_1), \ldots, (x_m, y_m)$ is also drawn i.i.d. from D_{XY}; the learning algorithm is permitted direct access to the sequence of x_i values (unlabeled data points), but has to make a label request to obtain the label y_i of example x_i. The hope is that we can output a classifier of small error by using many fewer label requests than in passive learning by actively directing the queries to informative examples (while keeping the number of unlabeled examples polynomial).

It has been long known that, in the realizable case, active learning can sometimes provide an exponential improvement in label complexity over passive learning. The canonical example [6] is learning threshold classifiers ($X = [0, 1]$ and $C = \{\mathbf{1}_{[0,a]} \mid a \in [0, 1]\}$). Here we can actively learn with only $\tilde{O}(\log(1/\epsilon))$ label requests by using a simple binary search-like algorithm as follows: we first draw $N = \tilde{O}((1/\epsilon)\log(1/\delta))$ unlabeled examples, then do binary search to find the transition from label 1 to label 0, and with only $O(\log(N))$ queries we can correctly infer the labels of all our examples; we finally output a classifier from C consistent with all the inferred labels. By standard VC-dimension based bounds for supervised learning [13], we are guaranteed to output an ϵ-accurate classifier. On the other hand, for passive learning, we provably need $\Omega(1/\epsilon)$ labels to output a classifier of error at most ϵ with constant probability, yielding the exponential reduction in label complexity.

Key Results

While in the simple threshold concept class described above active learning always provides huge improvements over passive learning, things are more delicate in more general scenarios. In particular, both in the realizable and in the agnostic case, it has been shown that for more general concept spaces, in the worst case over all data-generating distributions, the label complexity of active learning equals that of passive learning. Thus, much of the literature was focused on identifying non-worst case, natural conditions about the relationship between the data distribution and the target, under which active learning provides improvements over passive. Below, we discuss three approaches, under which active learning has been shown to reduce the label complexity: disagreement-based techniques, margin-based techniques and cluster-based techniques.

Disagreement-based active learning

Disagreement-based active learning was the first method to demonstrate the feasibility of *agnostic active learning* for general concept classes. The general algorithmic framework of disagreement-based active learning in the presence of noise was introduced with the A^2 algorithm by Balcan et al. [2]. Subsequently, several researchers have proposed related disagreement-based algorithms with improved sample complexity, *e.g.* [8, 11, 5].

At a high level, A^2 operates in rounds. It maintains a set of candidate classifiers from the concept class C and in each round queries labels aiming to efficiently reduce this set to only few high-quality candidates. More precisely, in round i, A^2 considers the set of surviving classifiers $C_i \subseteq C$, and asks for the labels of a few random points that fall in the *region of disagreement* of C_i. Formally, the region of disagreement of a set of classifiers C_i is $\mathrm{DIS}(C_i) = \{x \in X \mid \exists f, g \in C_i : f(x) \neq g(x)\}$. Based on these queried labels from $\mathrm{DIS}(C_i)$, to obtain C_{i+1}, the algorithm then throws out hypotheses that are suboptimal. The key ingredient is that A^2 *only* throws out hypotheses, for which it is *statistically confident* that they are suboptimal.

[2] show that A^2 provides exponential improvements in the label sample complexity in terms of the $1/\epsilon$-parameter when the noise rate η is sufficiently small, both for learning thresholds and for learning homogeneous linear separators in R^d, one of the most widely used and studied classes in machine learning. Following up on this, Hanneke [9] provided a generic analysis of the A^2 algorithm that applies to *any concept class.* This analysis quantifies the label complexity of A^2 in terms of the so-called *disagreement coefficient* of the class C. The disagreement coefficient is a distribution-dependent sample complexity measure that quantifies how fast the region of disagreement of the set of classifiers at distance r of the optimal classifier collapses as a function r. In particular, [9] showed that the label complexity of the A^2 algorithm is $O(\theta^2(\frac{\nu^2}{\epsilon^2}+1)(d\log(1/\epsilon)+\log(1/\delta))\log(1/\epsilon))$, where ν is the best error rate of a classifier in C, d is the VC-dimension of C, and θ is the disagreement-coefficient. As an example, for homogeneous linear separators, we have $\theta = \theta(\sqrt{d})$ under uniform marginal over the unit ball. Here, the disagreement-based analysis yields a label complexity of $\tilde{O}(d^2\frac{\nu^2}{\epsilon^2}\log(1/\epsilon))$ in the agnostic case and $\tilde{O}(d^{3/2}\log(1/\epsilon))$ in the realizable case.

Margin-based active learning

While the disagreement-based active learning line of work provided the first general understanding of the sample complexity benefits with active learning for arbitrary concept classes, it suffers from two main drawbacks: (1) methods and analyses developed in this context are often suboptimal in terms of label complexity, since they take a conservative approach and query even points on which there is only a small amount of uncertainty, (2) the methods are computationally inefficient. *Margin-based* active learning is a technique that overcomes both the above drawbacks for learning homogeneous linear separators under log-concave distributions. The technique was first introduced by Balcan et al. [3] and further developed by Balcan et al. [4], and Awasthi et al. [1].

At a high level, like disagreement-based methods, the margin-based active learning algorithm operates in rounds, in which a number of labels are queried in some subspace of the domain and a set of candidate classifiers for the next round is identified. The crucial idea to reduce the label complexity is to design a *more aggressive querying strategy* by carefully choosing where to query instead of querying in all of the current disagreement region. Concretely, in round k the algorithm has a *current hypothesis* w_k, and the set of candidate classifiers for the next round consists of all homogeneous halfspaces that lie in a *ball of radius* r_k *around* w_k (in terms of their angle with w_k). The algorithm then queries points for labels near the decision boundary of w_k; that is, it only queries points that are within a *margin* γ_k of w_k; see Figure 1. To obtain w_{k+1}, the algorithm finds a loss minimizer among the current set of candidates with respect to the queried examples of round k. In the realizable case, this is done by 0/1-loss minimization. In the presence of noise, to obtain a computationally efficient procedure, the margin-based technique minimizes a convex surrogate loss.

[3, 4] showed that by localizing aggressively, namely by setting the margin parameter to $\gamma_k = \Theta(\frac{1}{2^k})$, one can actively learn with only $\tilde{O}(d\log(1/\epsilon))$ label requests in the realizable case,

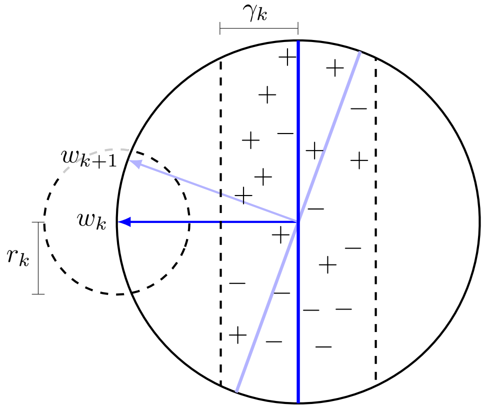

Figure 1: The margin-based active learning algorithm after iteration k. The algorithm samples points within margin γ_k of the current weight vector w_k and then minimizes the hinge loss over this sample subject to the constraint that the new weight vector w_{k+1} is within distance r_k from w_k.

when the underlying distribution is isotropic log-concave. A key idea of their analysis is to decompose, in round k, the error of a candidate classifier w as its error outside margin γ_k of the current separator plus its error inside margin γ_k, and to prove that for the above parameters, a small constant error inside the margin suffices to reduce the overall error by a constant factor. For the constant error inside the margin only $\theta(d)$ labels need to be queried, and since in each round the overall error gets reduced by a constant factor, $O(\log(1/\epsilon))$ rounds suffice to reduce the error to ϵ, yielding the label complexity of $\tilde{O}(d \log(1/\epsilon))$. Passive learning here provably requires $\Omega(d/\epsilon)$ labeled examples. Thus, the dependence on $1/\epsilon$ is exponentially improved, but without increasing the dependence on d (as in the disagreement-based method for this case, see above).

Building on this work, [1] gave the first *polynomial-time* active learning algorithm for learning linear separators to error ϵ in the presence of agnostic noise (of rate $O(\epsilon)$) when the underlying distribution is an isotropic log-concave distribution in R^d. They proposed to use a normalized hinge loss minimization (with normalization factor τ_k) for selecting the next classifier w_{k+1} in round k. [1] show that by setting the parameters appropriately (namely, $\tau_k = \Theta(1/2^k)$ and $r_k = \Theta(1/2^k)$), the algorithm again achieves error ϵ using only $O(\log(1/\epsilon))$ rounds, with $O(d^2)$ label requests per round. This yields a query complexity of $\mathrm{poly}(d, \log 1/\epsilon)$. The key ingredient for the analysis of this computationally efficient version in the noisy setting is proving that by constraining the search for w_{k+1} to vectors within a ball of radius r_k around w_k, the hinge-loss acts as a sufficiently faithful proxy for the 0/1-loss.

A recent work [14] proposes an elegant generalization of [3, 4] to more general concept spaces and shows an analysis that is always tighter than disagreement-based active learning (though their results are not computationally efficient).

Cluster-based active learning

The methods described above (disagreement-based and margin-based active learning) use active label queries to efficiently identify a classifier from the concept class C with low error. An alternative approach to agnostic active learning is to design active querying methods that efficiently find a (approximately) correct labeling of the unlabeled input sample. Here, “correct labeling” refers to the hidden labels y_i in the sample $(x_1, y_1), \ldots, (x_m, y_m)$ from the distribution D_{XY} (as defined in the formal setup section). The so labeled sample can then be used as input to a passive learning algorithm to learn an arbitrary concept class.

Cluster-based active learning is a method for the latter approach and was introduced by Dasgupta and Hsu [7]. The idea is to use a hierarchical clustering (cluster tree) of the unlabeled data, and check the clusters for label homogeneity by starting at the root of the tree (the whole data set) and working towards the leaves (single data points). The label homogeneity of a cluster is estimated by choosing data points for label query uniformly at random from the cluster. If a cluster is considered label homogeneous (with sufficiently high confidence), all remaining unlabeled points in that cluster are labeled with the majority label. If a cluster is detected to be label heterogeneous, it is split into its children in the cluster tree and processed later. The key insight in [7] is that since the cluster tree is fixed before any labels were seen, the induced labeled subsample of a child cluster can be considered a sample that was chosen uniformly at random from the points in that child-cluster. Thus, the algorithm can reuse labels from the parent cluster without introducing any sampling bias. The label efficiency of this paradigm crucially depends on the quality of input hierarchical clustering. Intuitively, if the cluster tree has a small pruning with label homogeneous clusters, the procedure will make only few label queries.

Urner et al. [12] proved label complexity reductions with this paradigm under a distributional assumption. They analyze a version (PLAL) of the above paradigm that uses hierarchical clusterings induced by *spatial trees* on the domain $[0,1]^d$ and provide label query bounds in terms of the *Probabilistic Lipschitzness* of the underlying data-generating distribution. Probabilistic Lipschitzness quantifies a marginal-label relatedness in the sense of close points being likely to have the same label. For a distribution with deterministic labels ($\mathbb{P}[Y = 1 \mid X = x] \in \{0, 1\}$ for all x), the Probabilistic Lipschitzness is a function ϕ that bounds, as a function of λ, the mass of points x for which both labels 0 and 1 occur in the ball $B_\lambda(x)$.

[12] show that, independently of the any data assumptions, (with probability $1 - \delta$) PLAL labels a $(1 - \epsilon)$-fraction of the input points correctly. They further show that using PLAL as a preprocedure, if the data-generating distribution has deterministic labels and its Probabilistic Lipschitzness is bounded by $\phi(\lambda) = \lambda^n$ for some $n \in \mathbb{N}$, then classes C of bounded VC-dimension on domain $X = [0,1]^d$ can be learned with $\tilde{O}((\frac{1}{\epsilon})^{\frac{n+2d}{n+d}})$ many labels, while any passive proper learner (*i.e.*, a passive learner that outputs a function from C) requires to see $\Omega(1/\epsilon^2)$ many labels. Further, [12] show that PLAL can be used to reduce the number of labels needed for nearest neighbor classification (*i.e.*, labeling a test point by the label of its nearest point in the sample) from $\Omega((\frac{1}{\epsilon})^{1+\frac{d-1}{n}})$ to $\tilde{O}((\frac{1}{\epsilon})^{1+\frac{d^2}{n(n+d)}})$.

Cross-References

PAC learning
Sample complexity
Computational complexity of learning

References

[1] P. Awasthi, M.-F. Balcan, and P. M. Long. The power of localization for efficiently learning linear separators with noise. In *Proceedings of the 46th Annual Symposium on the Theory of Computing (STOC), New York*, 2014.

[2] M. F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML), Pittsburgh*, 2006.

[3] M.-F. Balcan, A. Broder, and T. Zhang. Margin based active learning. In *Proceedings of the 20th Annual Conference on Computational Learning Theory (COLT), San Diego*, 2007.

[4] M.-F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of the 26th Conference on Learning Theory (COLT), Princeton*, 2013.

[5] A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang. Agnostic active learning without constraints. In *Advances in Neural Information Processing Systems (NIPS), Vancouver*, 2010.

[6] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. In *Proceedings of the 11th International Conference on Machine Learning (ICML), New Brunswick*, 1994.

[7] S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML), Helsinki*, 2008.

[8] S. Dasgupta, D.J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems (NIPS), Vancouver*, 2007.

[9] S. Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning (ICML), Corvallis*, 2007.

[10] Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press, Cambridge, MA, USA, 1994.

[11] V. Koltchinskii. Rademacher complexities and bounding the excess risk in active learning. *Journal of Machine Learning*, 11:2457–2485, 2010.

[12] R. Urner, S. Wulff, and S. Ben-David. Plal: Cluster-based active learning. In *Proceedings of the 26th Conference on Learning Theory (COLT), Princeton*, 2013.

[13] V. N. Vapnik. *Statistical Learning Theory.* John Wiley and Sons, New York, 1998.

[14] C. Zhang and K. Chaudhuri. Beyond disagreement-based agnostic active learning. In *Advances in Neural Information Processing Systems (NIPS), Montreal*, 2014.