Efficient Algorithms for General Active Learning

Open Problem, COLT '06

Claire Monteleoni

MIT

Is active learning a useful model?

Does AL (the PAC-like selective sampling) model help?

$\rightarrow$ By help we mean: yield label-complexity savings beyond PAC sample complexity.

Pose the simplest problem such that if AL is a useful model, it should be solvable.

$\rightarrow$ By useful we mean: studying the model yields (efficient) algorithms (with label-complexity bounds less than PAC).

To simplify problem: remove what could be solved via *unsupervised* learning.

- pinpoint AL problem only, to determine difficulty.

PAC-like selective sampling framework

Selective sampling [CAL92]:

Given: pool (or stream) of unlabeled examples, $x \in X$, drawn i.i.d. from input distribution, D over X.

Learner may request labels on examples in the pool/stream.

(Noiseless) oracle access to correct labels, $y \in Y$.

Constant cost per label

The error of any classifier v is measured on distribution D:

$\mathrm{err}(h) = P_{x\sim D}[v(x) \neq y]$

Goal: minimize label-complexity to learn the concept to a fixed error ε.

Non-Bayesian model: no prior on hypotheses assumed.

Open problem: efficient, general AL

Efficient algorithms for active learning under general input distributions, *D*.

$\rightarrow$ Current label-complexity upper bounds for general distributions are based on *intractable* schemes!

Provide an algorithm such that w.h.p.:

1. After *L* label queries, algorithm's hypothesis *v* obeys:

   $P_{x \sim D}[v(x) \neq u(x)] < \varepsilon$.

2. *L* is at most the PAC sample complexity, and for a general class of input distributions, *L* is significantly lower.
3. Running time is at most *poly*(d, 1/ε).

u is target, or best in class.

Open problem: specific variant

Efficient algorithms for active learning under general input distributions, *D*.

Specific variant: homogeneous linear separators, realizable case, *D* known to learner.

$$S = \left\{ x \in \mathbb{R}^d \;\middle|\; \|x\| = 1 \right\}, \quad x_t \in S, \quad y_t \in \{-1, +1\}$$

There exists a target $u : y_t(u \cdot x_t) > 0 \;\; \forall t, \quad \|u\| = 1$

D known:

Approximately, via an initial unsupervised learning phase, or

Exactly, in a new model:

Infinite unlabeled data for computing *D;*

Only have oracle access to labels on a finite subset

(cf. semi-supervised).

Open problem: specific variant

Efficient algorithms for active learning under general input distributions, *D*.

Specific variant: homogeneous linear separators, realizable case, *D* known to learner.

Standard PAC bound: $\tilde{O}(d/\varepsilon \log 1/\varepsilon)$.

Lower bound on label-complexity: $\Omega(1/\varepsilon)$ [D04].

$\rightarrow$ However, a pathological distribution yields bound.

If distribution is uniform: PAC complexity: $\Theta(d/\varepsilon)$ [L95,L03].

Label-complexity: $\tilde{O}(d \log 1/\varepsilon)$ [DKM05].

$\rightarrow$ What is a suitably "general class of input distributions"?

Open problem: other open variants

Efficient algorithms for active learning under general input distributions, *D*.

Other open variants:

- Input distribution, *D*, is unknown to learner.
- Agnostic case, certain scenarios.
- Add the online constraint: memory and time complexity (of the online update) must not scale with number of seen labels or mistakes.
- Same goal, other concept classes, or a general concept learner.

Related work: theory

Negative results:

Homogenous linear separators under arbitrary distributions and non-homogeneous under uniform: $\Omega(1/\varepsilon)$ [D04].

Perceptron algorithm under any AL rule uses $\Omega(1/\varepsilon^2)$ [DKM05].

Arbitrary (concept, distribution)-pairs that are "ρ-splittable": $\Omega(1/\rho)$ [D05].

Agnostic setting where best in class has generalization error β: $\Omega(\beta^2/\varepsilon^2)$ [K05].

Upper bounds on label-complexity not yet shown achievable by an (efficient) algorithm:

General concepts and input distributions, realizable:

e.g. $\tilde{O}(\log(1/\lambda)\, d \log^2(1/\varepsilon))$ for linear separators, under λ–similar to uniform [D05]. $\lambda \leq U(A)/P_D(A) \leq 1/\lambda \quad \forall A \subseteq X$

Linear separators under uniform, an agnostic scenario:

$\tilde{O}(d^2 \log 1/\varepsilon)$ [BBL06].

Related work: algorithms

Algorithms analyzed in other frameworks:

Individual sequence prediction, regret analysis: [C-BGZ05].

Bayesian assumption: linear separators, realizable case, using QBC algorithm [SOS92], label-complexity upper bounds: Uniform $\tilde{O}(d \log 1/\varepsilon)$ [FSST97].

λ–similar to uniform $\tilde{O}((1/\lambda)\, d \log 1/\varepsilon)$ [FSST97].

Label-complexity upper bounds when the input distribution is uniform:

Linear separators, realizable case, $\tilde{O}(d \log 1/\varepsilon)$ [DKM05].

Linear separators, realizable case, using [CAL92]. algorithm, $\tilde{O}(d^2 \log 1/\varepsilon)$ [BBL06].

Linear separators, realizable case, λ–similar to uniform, using [DKM05] algorithm, $\tilde{O}(\mathit{poly}(1/\lambda)\, d \log 1/\varepsilon)$ [M].

Thank you!

And thanks to:

Sanjoy Dasgupta

Matti Kääriäinen