10-806 Foundations of Machine Learning and Data Science

Lecturer: Maria-Florina Balcan Lecture 20 & 21: November 16 & 18, 2015

---

1 Active Learning

Most classic machine learning methods and the formal learning theory models we discussed (PAC [6] and Statistical Learning Theory [7]) 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 where the goal is to predict a binary label y based on its corresponding input vector x. We assume that the data points (x, y) are drawn from an unknown underlying distribution D_{XY} over $X \times Y$ where X is the 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 $err(h) = \Pr_{(x,y)\sim D_{XY}}[h(x) \neq y]$. As usual, we use D to denote the marginal distribution over X of D_{XY}. In *passive learning*, 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 (PAC learning), we assume that the true label of any example is determined by a deterministic function of the features (the target function) that belongs to a known concept class C. In the agnostic case, 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.

In *active learning*, a set of labeled examples $(x_1, y_1), \ldots, (x_m, y_m)$ is also drawn i.i.d. from D_{XY}. However, the learning algorithm is permitted direct access only to the sequence of x_i values (unlabeled data points). It must 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). The number of label requests made is called the *label complexity* of the algorithm.

Within active learning, we can further distinguish two models. In *pool-based* active learning, the algorithm is given as input a pool of unlabeled examples x_i and the algorithm can then query for the labels of examples of its choice from the pool in any order. In *selective sampling* active learning, the algorithm visits the unlabeled data points x_i in sequence, and, for each i, makes a decision on whether or not to request the label y_i based only on the previously-observed x_j values ($j \leq i$) and corresponding requested labels, and never changes this decision once made. In both cases, data is drawn from the distribution D_{XY}.

Let h^* denote the hypothesis of lowest error in C. So in the realizable case, $err_D(h^*) = 0$.

1.1 Disagreement coefficient

The disagreement coefficient is a measure of the complexity of an active learning problem that has proven quite useful for analyzing the certain types of active learning algorithms, the so called disagreement based active learning algorithms. It was introduced by Steve Hanneke in [4] in order to analyze the A^2 algorithm of Balcan, Beygelzimer, Langford [1] and it has been since the major notion of complexity of an active learning problem. It is also related to other measure of capacity in empirical processes, in particular Alexander capacity [5].

Informally, the disagreement coefficient quantifies how much disagreement there is among a set of classifiers relative to how close to h^* they are.

Definition 1. *Given some class C and some $V \subseteq C$, define:*

$$DIS(V) = \{x \in X : \exists h_1, h_2 \in V \text{ s.t. } h_1(x) \neq h_2(x)\}$$

and define

$$\Delta(V) = \Pr(x \in DIS(V)).$$

So, $\Delta(V)$ measures the probability that for a random instance x there will be two hypotheses in V that disagree on its label. For $r \in [0, 1]$ define:

$$\mathbf{B}(h, r) = \{h' \in C : d(h, h') \leq r\},$$

where

$$d(h, h') = \Pr_{x \sim D}(h(x) \neq h'(x)).$$

For example, in the realizable case, $\mathbf{B}(h^*, r)$ consists of all hypotheses in C of error rate at most r. Then $DIS(\mathbf{B}(h^*, r))$ is the set of all examples $x \in X$ such that at least two hypotheses of error rate at most r disagree on x.

Definition 2. *The disagreement coefficient of h^* with respect to C under distribution D:*

$$\Theta = \sup_{r>0} \frac{\Pr(DIS(\mathbf{B}(h^*, r)))}{r} = \sup_{r>0} \frac{\Delta(\mathbf{B}(h^*, r))}{r}.$$

Examples

Thresholds Let's assume that the instance space is $X = [0, 1]$ and let $h_z(x) = 1$ if $x \geq z$ and $h_z(x) = -1$ otherwise. For simplicity let's assume that D is uniform in the interval $[0, 1]$. Then $DIS(\mathbf{B}(h_z, r)) = [z - r, z + r]$. So $\Pr(DIS(\mathbf{B}(h_z, r))) = 2r$. This means that the disagreement coefficient $\Theta_{h_z} = 2$ for all h_z. A similar argument holds for any underlying distribution D.

Intervals Let's assume that the instance space is $X = [0, 1]$, and let $h_{[a,b]}$ be the hypothesis that labels examples as positive iff they fall in the interval $[a, b]$. For simplicity let's assume that D is uniform in the interval $[0, 1]$. Then we have $\Theta_{h_{[a,b]}} = \max(\frac{1}{b-a}, 4)$. Specifically, in the case that $r < b - a$ we have $\Pr(DIS(\mathbf{B}(h_{[a,b]}, r))) = 4r$ using the same reasoning as for the case of thresholds on the line. For the case that $r \geq b - a$ we have that every interval of width $\leq r - (b - a)$ is in

$\mathbf{B}(h_{[a,b]}, r)$. This means that $\Pr(DIS(\mathbf{B}(h_{[a,b]}, r))) = 1$. Combining the two cases together we get $\Theta_{h_{[a,b]}} = \max(\frac{1}{b-a}, 4)$.

Linear separators under nice distributions. Let C be the class of homogeneous linear separators in R^d and let's assume that D is an isotropic log-concave distribution in R^d. Then the disagreement coefficient is $O(\sqrt{d})$ (See [2] for a proof.)

Intuitively, for small r we expect $\Delta(\mathbf{B}(h, r))$ to become smaller. The disagreement coefficient measures how quickly $\Delta(\mathbf{B}(h, r))$ grows/shrinks as a function of r.

1.2 The CAL algorithm

The CAL algorithm is for active learning in the realizable case [3][1] and proceeds as follows.

Algorithm 1 CAL

Input: parameters k and ϵ.

1. Begin with $V = C$.
2. While $\Delta(V) > \epsilon$ do: (each run through this loop is one round)
   (a) Keep sampling from D until one has collected k instances in $DIS(V)$. Notice that determining membership in $DIS(V)$ can be done without observing any labels.
   (b) Query for the labels of the k examples in $DIS(V)$. Call them $(x_1, y_1), \ldots, (x_k, y_k)$.
   (c) Update: $V \leftarrow \{h \in V : \forall\ j \in [k], h(x_j) = y_j\}$. I.e., V is the current *version space*, namely the set of hypotheses consistent with all labeled examples so far.

Output: Any hypothesis from V.

Theorem 1. *Assume Θ is finite and let d=VCdim(C). Then for any given ϵ, δ, if Algorithm 1 is run with $k = \tilde{O}(\Theta d \ln \Theta)$, it will output a hypothesis with error $\leq \epsilon$ with probability $\geq 1 - \delta$, and it will stop after $O(\log 1/\epsilon)$ rounds. The label complexity is $\tilde{O}(\Theta \log(1/\epsilon) d \ln \Theta)$.*

Proof. Let V_i be the version space at round i. First of all, if $\Delta(V_i) \leq \epsilon$, then since we never eliminate h^*, all the hypotheses in V_i have error $\leq \epsilon$.

To bound the label complexity, we show that $\Delta(V_{i+1}) \leq \frac{1}{2}\Delta(V_i)$ with high probability; the overall label complexity bound then follows from the definition of k.

Let

$$V_i^{\Theta} = \left\{ h \in V_i : err(h) = d(h, h^*) \geq \frac{\Delta(V_i)}{2\Theta} \right\},$$

i.e., the hypotheses in V_i with large error.

We will show that after k examples, we have $V_{i+1} \subseteq V_i \setminus V_i^{\Theta}$ whp. Let's assume this for now. In that case, since $V_i \setminus V_i^{\Theta} \subseteq \mathbf{B}(h^*, \frac{\Delta(V_i)}{2\Theta})$ we have

$$\Delta(V_{i+1}) \leq \Delta\left(\mathbf{B}\left(h^*, \frac{\Delta(V_i)}{2\Theta}\right)\right) \leq \Theta \frac{\Delta(V_i)}{2\Theta} = \frac{\Delta(V_i)}{2}$$

[1] The original paper [3] presents the algorithm (with motivation and experiments), but provides no formal analysis. A scheme of this type was first analyzed in [1] and later generalized by many others (see [5]).

as desired.

Now, we show that $V_{i+1} \subseteq V_i \setminus V_i^{\Theta}$ after k examples whp. Let D_i denote the conditional distribution of D given that $x \in DIS(V_i)$. By definition, for all $h \in V_i$,

$$\Delta(V_i) err_{D_i}(h) = err_D(h) = d(h, h^*),$$

because outside of $DIS(V_i)$ we know that h and h^* agree.

If $h \in V_i^{\Theta}$ we have

$$d(h, h^*) \geq \frac{\Delta(V_i)}{2\Theta},$$

$$\Delta(V_i) err_{D_i}(h) \geq \frac{\Delta(V_i)}{2\Theta},$$

$$err_{D_i}(h) \geq \frac{1}{2\Theta}.$$

The key point now is that by standard sample complexity bounds, we only need to sample $k = \tilde{O}(\Theta d \ln \Theta)$ points from D_i to eliminate all such hypotheses whp, where the $\tilde{O}$ is hiding the $\log(N/\delta)$ term where N is the number of rounds, so that we have a δ failure probability overall by the union bound.

So, with $\tilde{O}(\Theta d \ln \Theta)$ labeled examples we halve the region of disagreement as desired, finishing the proof. □

1.3 The agnostic case and the A^2 algorithm

We describe here a simplification of the A^2 algorithm due to [1] and its general analysis due to [4]

Let h^* now denote the function of lowest true error within class C.

Definition 3. *A subroutine for computing $LB(S, h, \delta)$ and $UB(S, h, \delta)$ is said to be legal if for all distributions D_{XY} over $X \times Y$, for all $0 < \delta < 1/2$, and all integers m, we have that with probability $\geq 1 - \delta$ over the draw of S from D_{XY}^m, for all $h \in C$, $LB(S, h, \delta) \leq err_P(h) \leq UB(S, h, \delta)$.*

E.g., based on the results from Lecture 7 (September 30th, 2015), for classes of VC-dimension d we have $UB(S, h, \delta) = \min\left[err_S(h) + \sqrt{c\frac{d\log(m/d)+\log(1/\delta)}{m}}, 1\right]$ and
$LB(S, h, \delta) = \max\left[err_S(h) - \sqrt{c\frac{d\log(m/d)+\log(1/\delta)}{m}}, 0\right]$.

We now present Algorithm2, a simplified version of the A^2 algorithm for which the analysis is much cleaner.

Theorem 2. *Assume $\nu = err_D(h^*)$. If Algorithm 2 is run with parameters ν, Θ, δ and our usual VC-dimension upper and lower bounds, then with probability at least $1 - \delta$, it outputs a hypothesis of error at most $\nu + \epsilon$, and will query at most*

$$\tilde{O}\left(\Theta^2 d \frac{\nu^2}{\epsilon^2} \log\left(\frac{\Theta\nu}{\epsilon}\right)\right)$$

labels. (supposing $\epsilon \leq \nu$ for simplicity)

Algorithm 2 A^2

Input: Parameters ν, Θ, δ.

1. Initialize $V_i = C$, $k = \tilde{O}(\Theta^2 d)$, $k' = \tilde{O}(\Theta^2 d\nu^2/\epsilon^2)$, $\delta' = \delta/(1 + \lceil \log(\frac{1}{8\Theta\nu}) \rceil)$.
2. While $\Delta(V_i) \geq 8\Theta\nu$ do:
   (a) Let D_i be the conditional distribution D given that $x \in DIS(V_i)$.
   (b) Sample k iid labeled examples from D_i. Denote this set by S_i.
   (c) Update $V_{i+1} = \{h \in V_i : LB(S_i, h, \delta') \leq \min_{h' \in H} UB(S_i, h', \delta')\}$.

Output: Sample S of k' points from D_i and output $\text{argmin}_{h \in V_i} err_S(h)$.

Proof. Assume first that $\Delta(V_i) \leq 8\Theta\nu$ and $h^* \in V_i$. For all $h \in V_i$ we have

$$err_D(h) - err_D(h^*) = \Delta(V_i)(err_{D_i}(h) - err_{D_i}(h^*)) \leq 8\Theta\nu(err_{D_i}(h) - err_{D_i}(h^*)).$$

So, to find an $h \in V_i$ with $err_D(h) \leq err_D(h^*) + \epsilon$ it suffices to find $h \in V_i$ with

$$err_{D_i}(h) - err_{D_i}(h^*) \leq \frac{\epsilon}{8\Theta\nu}.$$

Standard error bounds tell us that $O(k')$ examples from D_i are sufficient to guarantee that ERM over V_i will find an h such that $err_{D_i}(h) - err_{D_i}(h^*) \leq \frac{\epsilon}{8\Theta\nu}$ as desired.

At this point we have analyzed the very last step. What remains is to analyze how many rounds we need to reach V_i with $\Delta(V_i) \leq 8\Theta\nu$, and also to argue that whp h^* is not thrown out. Specifically, we will argue that whp after each round we have $\Delta(V_{i+1}) \leq \Delta(V_i)/2$, so that the total number of rounds is at most $\lceil \log(\frac{1}{8\Theta\nu}) \rceil + 1$, and h^* is not removed.

Define

$$V_i^{\Theta} = \left\{ h \in V_i : d(h, h^*) \geq \frac{\Delta(V_i)}{2\Theta} \right\}.$$

We now aim to show that with high probability $V_{i+1} \subseteq V_i \setminus V_i^{\Theta}$. Since $V_i \setminus V_i^{\Theta} \subseteq \mathbf{B}(h^*, \frac{\Delta(V_i)}{2\Theta})$, as in the analysis of CAL this would then imply that $\Delta(V_{i+1}) \leq \Delta(V_i)/2$ by definition of the disagreement coefficient.

First of all, for $h \in V_i$ we have:

$$d(h, h^*) \leq \Delta(V_i) \Pr_{x \sim D_i}(h(x) \neq h^*(x)) \leq \Delta(V_i)[err_{D_i}(h) + err_{D_i}(h^*)] \leq \Delta(V_i) err_{D_i}(h) + \nu.$$

Now, if $d(h, h^*) \geq \frac{\Delta(V_i)}{2\Theta}$ then by the above we have

$$err_{D_i}(h) \geq \frac{1}{2\Theta} - \frac{\nu}{\Delta(V_i)} \geq \frac{1}{2\Theta} - \frac{1}{8\Theta} = \frac{3}{8\Theta}.$$

On the other hand we have

$$err_{D_i}(h^*) \leq \frac{\nu}{\Delta(V_i)} \leq \frac{1}{8\Theta}.$$

So,

$$err_{D_i}(h) - \frac{1}{8\Theta} \geq err_{D_i}(h^*) + \frac{1}{8\Theta}.$$

This means that by the standard sample complexity bounds it is enough to draw $\tilde{O}(\Theta^2 d \ln(\Theta))$ examples to whp remove all $h \in V_i^{\Theta}$ as desired.

Finally, whp h^* is never removed from V_i because no hypothesis can have an upper bound lower than the lower bound for h^*. □

For a much more comprehensive description of disagreement based active learning see the Hanneke survey [5].

References

[1] M. F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *ICML*, 2006.

[2] M. F. Balcan and P. M. Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, 2013.

[3] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2), 1994.

[4] S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007.

[5] S. Hanneke. *Theory of Disagreement-Based Active Learning.* Foundations and Trends in Machine Learning, 2014.

[6] M. Kearns and U. Vazirani. *An introduction to computational learning theory.* MIT Press, Cambridge, MA, 1994.

[7] V. Vapnik. *Statistical Learning Theory.* Wiley-Interscience, 1998.