Advanced Course in Machine Learning Spring 2010

Active Learning: Disagreement Coefficient

Handouts are jointly prepared by Shie Mannor and Shai Shalev-Shwartz

In previous lectures we saw examples in which active learning gives an exponential improvement in the number of labels required for learning. In this lecture we describe the *Disagreement Coefficient* —a measure of the complexity of an active learning problem proposed by Steve Hanneke in 2007. We will derive an algorithm for the realizable case and analyze it using the disagreement coefficient. In particular, we will show that if the disagreement coefficient is constant then it is possible to obtain exponential improvement over passive learning. We will also describe a variant of the Agnostic Active (A^2) algorithm (due to Balcan, Beygelzimer, Langford) and show how an exponential improvement can be obtained even in the agnostic case, as long as the accuracy is proportional to the best error rate.

1 Motivation

Recall that for the task of learning thresholds on the line we established a logarithmic improvement for active learning using a binary search. Now, lets consider a different algorithm that achieves the same label complexity bound but is less specific.

For simplicity, assume that $\mathcal{D}_x$ is uniform over $[0,1]$. We start with $V_1 = [0,1]$, which represents the initial version space. Now we sample a random example (x_1, y_1) and update $V_2 = \{a \in V_1 : \text{sign}(x_1 - a) = y_1\}$. That is, if $y_1 = 1$ then now $V_2 = [0, x_1]$ and otherwise $V_2 = [x_1, 1]$. Next, we sample instances until we have $x_2 \in V_2$. We query the label y_2 and again update $V_3 = \{a \in V_2 : \text{sign}(x_2 - a) = y_2\}$. We continue this process until $V_i = [a_i, b_i]$ satisfies $b_i - a_i \le \epsilon$.

Lets analyze this algorithm. First, it is easy to verify that if we stop then all hypotheses in V_i has error of at most ϵ (recall that we assume the data is realizable therefore $a^\star \in V_i$ at all times). Second, whenever we query the label we have that x_i is selected randomly from $V_i = [a_i, b_i]$. Denote $\Delta(V_i) = b_i - a_i$. Whenever we query $x_i \in [a_i, b_i]$, if $x_i \in [a_i + \Delta(V_i)/4, b_i - \Delta(V_i)/4]$ then the size of V_{i+1} satisfies $\Delta(V_{i+1}) \le (3/4)\Delta(V_i)$. The probability that $x_i \in [a_i + \Delta(V_i)/4, b_i - \Delta(V_i)/4]$ is $1/2$ so the probability that this will not happen in k consecutive trials is 2^{-k}. For $n = kr$, we can rewrite $\Delta(V_{n+1})$ as

$$\Delta(V_{n+1}) = \Delta(V_1) \left(\frac{\Delta(V_2)}{\Delta(V_1)} \cdots \frac{\Delta(V_{k+1})}{\Delta(V_k)} \right) \cdots \left(\frac{\Delta(V_{n-k})}{\Delta(V_{n-k-1})} \cdots \frac{\Delta(V_{n+1})}{\Delta(V_n)} \right)$$

All the multiples above are at most 1. Additionally, for each k multiples in each parenthesis, with probability of $1-2^{-k}$ at least one multiple is at most $3/4$. Therefore, with probability of at least $1-n2^{-k}$ the expression within each parenthesis is at most $3/4$. It follows that with probability of at least $1 - n2^{-k}$ we have

$$\Delta(V_{n+1}) \le (3/4)^{n/k} .$$

To make the above at most ϵ it suffices that $n \ge \Omega(k \log(1/\epsilon))$ and to make the probability of failure be at most δ we can choose $k = \lceil \log_2(n/\delta) \rceil$. In summary, we have shown that with probability of at least $1 - \delta$, the label complexity to achieve accuracy of ϵ is

$$n = O\left(\log(n/\delta)\, \log(1/\epsilon)\right) .$$

The above analysis relied on the property that it is relatively easy to significantly decrease $\Delta(V_i)$ by sampling from V_i. In the next section we describe the disagreement coefficient which characterizes when such an approach will work.

2 The Disagreement Coefficient

Let $\mathcal{H}$ be a hypothesis class. We define a pseudo-metric on $\mathcal{H}$, based on the marginal distribution $\mathcal{D}_x$ over instances, such that

$$d(h,h') = \mathop{\mathbb{P}}_{x\sim\mathcal{D}_x}[h(x) \neq h'(x)] \ .$$

We define the corresponding ball of radius r around $h^\star$ to be

$$B(h^\star, r) = \{h \in \mathcal{H} : d(h,h^\star) \leq \epsilon\} \ .$$

The disagreement region of a subset of hypotheses $V \subset \mathcal{H}$ is

$$\mathrm{DIS}(V) = \{x : \exists h, h' \in V,\ h(x) \neq h'(x)\}$$

and its mass is

$$\Delta(V) = \mathcal{D}_x(\mathrm{DIS}(V)) = \mathbb{P}[x \in \mathrm{DIS}(V)] \ .$$

That is, $\Delta(V)$ measures the probability to choose an instance x such that there are two hypotheses in V that disagree on its label. Intuitively, for small r we expect $\Delta(B(h^\star,r))$ to become smaller and smaller. The *disagreement coefficient* measures how quickly $\Delta(B(h^\star,r))$ grows as a function of r.

Definition 1 (Disagreement Coefficient) *Let $\mathcal{H}$ be a hypothesis class, $\mathcal{D}$ be a distribution over $\mathcal{X} \times \{0,1\}$, and $\mathcal{D}_x$ be the marginal distribution over $\mathcal{X}$. Let $h^\star$ be a minimizer of* $\mathrm{err}_{\mathcal{D}}(h)$*. The disagreement coefficient is*

$$\theta \stackrel{\text{def}}{=} \sup_{r\in(0,1)} \frac{\Delta(B(h^\star,r))}{r} \ .$$

The disagreement coefficient allows us to bound the mass of the disagreement region of a ball by its radius since

$$\forall r \in (0,1),\ \ \Delta(B(h^\star,r)) \leq \theta\, r \ .$$

2.1 Examples

Thresholds on the line: Let $\mathcal{H}$ be threshold functions of the form $x \mapsto \mathrm{sign}(x-a)$ for $a \in \mathbb{R}$ and $\mathcal{X} = \mathbb{R}$. Take two hypotheses h, h' in $B(h^\star, r)$. Denote $a, a', a^\star$ the thresholds of $h, h', h^\star$ respectively. Let a_0 be the minimal threshold such that $\mathcal{D}_x([a_0, a^\star]) \leq r$ and similarly let a_1 be the maximal threshold such that $\mathcal{D}_x([a^\star, a_1]) \leq r$. Clearly, both a and a' are in $[a_0, a_1]$. It follows that $\Delta(B(h^\star,r)) \leq 2r$ and therefore $\theta \leq 2$.

Halfspaces: TBA

Finite classes: Let $\mathcal{H} = \{h_1, \ldots, h_k\}$ be a finite class. We show that $\theta \leq |\mathcal{H}|$ and that there are classes for which equality holds. The upper bound is derived as follows. First, for each class of the form $\{h_i, h^\star\}$ we have $\theta = 1$. Second, it is easy to verify that the disagreement coefficient of $H_1 \cup H_2$ with respect to $h^\star \in H_1 \cap H_2$ is at most the sum of the disagreement coefficients. From this it follows that $\theta \leq |\mathcal{H}|$. Finally, an example in which $\theta = |\mathcal{H}|$ is a class in which for all i such that $h_i \neq h^\star$ we have that $h_i(x_i) \neq h^\star(x_i)$ and for all other instances h_i agrees with $h^\star$. In such a case, setting $r = 1/k$ gives the matching lower bound.

3 An algorithm for a finite class in the realizable case

To demonstrate how the disagreement coefficient affects the performance of active learning we start with a simple algorithm for the case of a finite $\mathcal{H}$ and assuming that there exists $h^\star \in \mathcal{H}$ with $\mathrm{err}_\mathcal{D}(h^\star) = 0$.

Algorithm 1 ActiveLearning

```
Parameters: $k, \epsilon$
Initialization $V = \mathcal{H}$
Loop while $(\Delta(V) > \epsilon)$
  Keep sampling instances until having $k$ instances in $\mathrm{DIS}(V)$
  Query their labels
  Let $(x_1, y_1), \ldots, (x_k, y_k)$ be the resulting examples
  Update: $V = \{h \in V : \forall j \in [k],\ h(x_j) = y_j\}$
end while
Output: Any hypothesis from $V$
```

We now analyze the label complexity of the above algorithm.

Theorem 1 *Let $\mathcal{H}$ be a finite hypothesis class, $\mathcal{D}$ be a distribution, and assume that the disagreement coefficient, θ, is finite. For any $\delta, \epsilon \in (0, 1)$, if Algorithm 1 is run with ϵ and with $k = \lceil 2\theta \log(n|\mathcal{H}|/\delta) \rceil$, where $n = \lceil \log_2(1/\epsilon) \rceil$, then the algorithm outputs a hypothesis with $\mathrm{err}_\mathcal{D}(h) \le \epsilon$ and with probability of at least $1 - \delta$ will stop after at most n rounds. The label complexity is therefore*

$$O\left(\theta\ \log(1/\epsilon)\ \left(\log(|\mathcal{H}|/\delta) + \log\log(1/\epsilon)\right)\right) .$$

Proof Let V_i be the version space at round i. First, if we stop then $\Delta(V_i) \le \epsilon$ and this guarantees that the error of all hypotheses in V_i is at most ϵ.

To bound the label complexity of the algorithm we will show that $\Delta(V_{i+1}) \le \frac{1}{2}\Delta(V_i)$ with high probability. Let $V_i^\theta = \{h \in V_i : d(h, h^\star) > \Delta(V_i)/(2\theta)\}$ be all hypotheses in V_i with a large error. We will show that $V_{i+1} \subseteq V_i \setminus V_i^\theta$ and since $V_i \setminus V_i^\theta \subseteq B(h^\star, \Delta(V_i)/(2\theta))$ we shall have

$$\Delta(V_{i+1}) \le \Delta(B(h^\star, \Delta(V_i)/(2\theta))) \le \theta\,\Delta(V_i)/(2\theta) = \Delta(V_i)/2 ,$$

where in the last inequality we used the definition of the disagreement coefficient.

Indeed, let $\mathcal{D}_i$ be the conditional distribution of $\mathcal{D}$ given that x in $\mathrm{DIS}(V_i)$ and note that in the realizable case we have

$$\Delta(V_i)\,\mathrm{err}_{\mathcal{D}_i}(h) = \mathrm{err}_\mathcal{D}(h) = d(h, h^\star) .$$

It follows that $h \in V_i^\theta$ iff $d(h, h^\star) > \Delta(V_i)/(2\theta)$ iff $\mathrm{err}_{D_i}(h) > 1/(2\theta)$. Therefore, for $h \in V_i^\theta$, the probability to choose $x \in \mathrm{DIS}(V_i)$ for which $h(x) \ne h^\star(x)$ is at least $1/(2\theta)$. Since we sample k points from $\mathrm{DIS}(V_i)$ we obtain that with probability of at least $1-(1-1/(2\theta))^k$ at least one of the k points satisfies $h(x) \ne h^\star(x)$ which means that h will not be in V_{i+1}. Applying the union bound over all hypotheses in V_i^θ and note that $V_i^\theta \subseteq \mathcal{H}$ we obtain

$$\mathbb{P}[\exists h \in V_{i+1} \cap V_i^\theta] \le |\mathcal{H}|\,(1 - 1/(2\theta))^k \le |\mathcal{H}| \exp(-k/(2\theta)) .$$

Choosing k as in the theorem statement we get that with probability of at least $1 - \delta/n$ we have that $V_{i+1} \subseteq V_i \setminus V_i^\theta$ and as we showed before this implies $\Delta(V_{i+1}) \le \frac{1}{2}\Delta(V_i)$. Applying a union bound over the first n rounds of the algorithm we obtain that with probability of at least $1 - \delta$,

$$\Delta(V_n) \le \Delta(V_1)2^{-n} \le 2^{-n} ,$$

which will be smaller than ϵ if $n = \lceil \log_2(1/\epsilon) \rceil$. ■

3.1 An extension to a low VC class

It is easy to tackle the case of a hypothesis class with VC dimension of d based on the above. The idea is to first sample m unlabeled examples. Based on Sauer lemma, the restriction of $\mathcal{H}$ to these m examples is of size at most $O(m^d)$. Now, apply the analysis from previous section for finite classes w.r.t. the uniform distribution over the sample. It gives an almost ERM for the m examples. The result follows from generalization bounds for almost ERMs.

3.2 Query by Committee revisited

Recall that the QBC algorithm query the label if two random hypotheses from V_i give a different answer. This is similar in spirit to sampling from the disagreement region. We left as an exercise to analyze QBC using the concept of disagreement coefficient.

4 The A^2 algorithm

In this section we describe and analyze the *Agnostic Active* (A^2) algorithm, originally proposed by Balcan, Beygelzimer and Langford in 2006. We give a variant of the algorithm which is easier to analyze. A more complicated variant of the algorithm is described and analyze in Hanneke (2007).

The algorithm is similar to Algorithm 1. We sample examples and query their labels only if they fall in the disagreement region. The major difference is that while in Algorithm 1, we throw away a hypothesis even if it makes a single error, now we throw away a hypothesis only if we are certain that it is worse than $h^\star$.

The algorithm below relies on a function $\mathrm{UB}(S,h)$, which provides an upper bound on $\mathrm{err}_{\mathcal{D}}(h)$ (one that holds with high probability), and $\mathrm{LB}(S,h)$, which provides a lower bound on $\mathrm{err}_{\mathcal{D}}(h)$. Such bounds can be obtained from VC theory. For simplicity, we focus on finite hypotheses class. In that case we have:

Lemma 1 *Let $\mathcal{H}$ be a finite hypothesis class, let $\mathcal{D}$ be a distribution, and let $S \sim \mathcal{D}^m$ be a training set of m examples sampled i.i.d. from $\mathcal{D}$. Then, with probability of at least $1-\delta$ we have*

$$\forall h \in \mathcal{H}, \;\; |\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| \le \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2m}} \; .$$

Based on the above lemma, we define

$$\mathrm{UB}(S,h) = \min\left\{\mathrm{err}_S(h) + \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2m}}, 1\right\} \quad ; \quad \mathrm{LB}(S,h) = \max\left\{\mathrm{err}_S(h) - \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2m}}, 0\right\}$$

and we clearly have that with probability of at least $1-\delta$

$$\forall h \in \mathcal{H}, \;\; \mathrm{LB}(S,h) \;\le\; \mathrm{err}_{\mathcal{D}}(h) \;\le\; \mathrm{UB}(S,h)$$

Algorithm 2 Agnostic Active (A^2)

```
Parameters: $\nu, \theta$, and $\delta$ (used in UB and LB)
Initialization: $V_1 = \mathcal{H}$ ; $k = \lceil 32\theta^2 \log(2|\mathcal{H}|/\delta) \rceil$
Loop while $\Delta(V_i) > 8\theta\nu$
  Let $\mathcal{D}_i$ be the conditional distribution $\mathcal{D}$ given that $x \in \mathrm{DIS}(V_i)$
  Sample $k$ i.i.d. labeled examples from $\mathcal{D}_i$, denote this set by $S_i$
  Update: $V_{i+1} = \{h \in V_i : \mathrm{LB}(S_i,h) \le \min_{h' \in \mathcal{H}} \mathrm{UB}(S_i,h')\}$
end while
Output: Sample $S \sim \mathcal{D}_i^{8k}$ and return $\mathrm{argmin}_{h \in V_i} \mathrm{err}_S(h)$
```

The following theorem shows that we can achieve a constant approximation of the best possible error very quickly.

Theorem 2 *Let $\mathcal{H}$ be a finite hypothesis class, $\mathcal{D}$ be a distribution, and assume that the disagreement coefficient, θ, is finite. Assume that $\nu \geq \mathrm{err}_{\mathcal{D}}(h^\star)$ for the optimal $h^\star \in \mathcal{H}$. Then, if Algorithm 2 runs with ν, θ, δ, then with probability of at least $1 - \delta \log(1/(8\theta\nu))$ the algorithm outputs a hypothesis with $\mathrm{err}_{\mathcal{D}}(h) \leq 2\nu$ and will query at most*

$$O\left(\theta^2 \log(1/(\theta\nu)) \log(|\mathcal{H}|/\delta)\right)$$

labels.

Proof First, suppose that $\Delta(V_i) \leq 8\theta\nu$ and that $h^\star \in V_i$. Then, since for all $h \in V_i$ we have

$$\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_{\mathcal{D}}(h^\star) = \Delta(V_i)(\mathrm{err}_{\mathcal{D}_i}(h) - \mathrm{err}_{\mathcal{D}_i}(h^\star)) \leq 8\theta\nu(\mathrm{err}_{\mathcal{D}_i}(h) - \mathrm{err}_{\mathcal{D}_i}(h^\star)) ,$$

it follows that to find $h \in V_i$ with $\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_{\mathcal{D}}(h^\star) \leq \nu$ it suffices to find $h \in V_i$ with $\mathrm{err}_{\mathcal{D}_i}(h) - \mathrm{err}_{\mathcal{D}_i}(h^\star) \leq 1/(8\theta)$. Standard generalization bounds tell us that taking an i.i.d. sample from $\mathcal{D}_i$ of size $4k$ guarantees that the ERM rule satisfies $\mathrm{err}_{\mathcal{D}_i}(h) - \mathrm{err}_{\mathcal{D}_i}(h^\star) \leq 1/(8\theta)$ with probability of at least $1 - \delta$. Hence, it is left to analyze how many rounds are required to have $\Delta(V_i) \leq 8\theta\nu$ and to verify that $h^\star \in V_i$ at all times.

Let $n = \lceil \log(1/(8\theta\nu)) \rceil + 1$ and let $\delta' = n\delta$. We will show that with probability of at least $1 - \delta'$ we have that $h^\star$ is never removed from V_i and that $\Delta(V_{i+1}) \leq \Delta(V_i)/2$ on all rounds from 1 to n. This will imply that $\Delta(V_n) \leq 8\theta\nu$ and therefore the algorithm will stop after at most n rounds.

Based on Lemma 1 we have that for all rounds from 1 to n and all h, with probability of at least $1 - \delta'$ we have that

$$|\mathrm{err}_{\mathcal{D}_i}(h) - \mathrm{err}_{S_i}(h)| \leq \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2k}} .$$

In particular this means we never remove $h^\star$ from V_i because no hypothesis can have a lower bound smaller then the upper bound for $h^\star$.

Similarly to the proof of Theorem 1 let $V_i^\theta = \{h \in V_i : d(h, h^\star) > \Delta(V_i)/(2\theta)\}$ be all hypotheses in V_i with a large disagreement with $h^\star$. We will show that $V_{i+1} \subseteq V_i \setminus V_i^\theta$ and since $V_i \setminus V_i^\theta \subseteq B(h^\star, \Delta(V_i)/(2\theta))$ we shall have

$$\Delta(V_{i+1}) \leq \Delta(B(h^\star, \Delta(V_i)/(2\theta))) \leq \theta\, \Delta(V_i)/(2\theta) = \Delta(V_i)/2 .$$

Note that for each $h \in V_i$ we have that if $h(x) \neq h^\star(x)$ then $x \in \mathrm{DIS}(V_i)$. Therefore,

$$d(h, h^\star) \leq \Delta(V_i) \mathop{\mathbb{P}}_{x \in \mathcal{D}_i}[h(x) \neq h^\star(x)] \leq \Delta(V_i)(\mathrm{err}_{D_i}(h) + \mathrm{err}_{D_i}(h^\star)) \leq \Delta(V_i)\mathrm{err}_{D_i}(h) + \nu ,$$

where the last inequality is because $\nu \geq \mathrm{err}_{\mathcal{D}}(h^\star) \geq \Delta(V_i)\mathrm{err}_{D_i}(h^\star)$. Therefore, if $d(h, h^\star) > \Delta(V_i)/(2\theta)$ then

$$\begin{aligned} \frac{\Delta(V_i)}{2\theta} &< d(h, h^\star) \leq \Delta(V_i)\mathrm{err}_{D_i}(h) + \nu \\ \Rightarrow \quad \mathrm{err}_{D_i}(h) &> \frac{1}{2\theta} - \frac{\nu}{\Delta(V_i)} \end{aligned}$$

Using again $\nu/\Delta(V_i) \geq \mathrm{err}_{\mathcal{D}_i}(h^\star)$ we get

$$\begin{aligned} \mathrm{err}_{D_i}(h) &> \frac{1}{2\theta} - \frac{\nu}{\Delta(V_i)} + \left(\mathrm{err}_{\mathcal{D}_i}(h^\star) - \frac{\nu}{\Delta(V_i)}\right) \\ \Rightarrow \quad \mathrm{err}_{D_i}(h) - \frac{1}{8\theta} &> \mathrm{err}_{\mathcal{D}_i}(h^\star) + \frac{3}{8\theta} - 2\frac{\nu}{\Delta(V_i)} \end{aligned}$$

Since we now assume that $\Delta(V_i) > 8\theta\nu$ the above implies

$$\mathrm{err}_{D_i}(h) - \frac{1}{8\theta} > \mathrm{err}_{\mathcal{D}_i}(h^\star) + \frac{1}{8\theta} .$$

Since $k \geq 32\theta^2 \log(2|\mathcal{H}|/\delta)$ then all hypotheses in V_i^θ will be removed. Thus, $\Delta(V_{i+1}) \leq \Delta(V_i)/2$.

■

4.1 An extension to a low VC class

It is easy to tackle the case of a hypothesis class with VC dimension of d using the same technique as in the previous section. Hanneke gave a direct analysis, without using the Sauer lemma trick