COMS 6998-4 Fall 2017 Presenter: Daniel Hsu
October 4, 2017 Scribe: Jiefu Ying

Active Learning Analysis in Agnostic Case

1 Introduction

Last lecture has introduced an active learning algorithm works in realizable case, while the assumption on the optimal hypothesis of set $\mathcal{H}$ may not hold in most cases. In other words, probably there isn't a "perfect" hypothesis included in $\mathcal{H}$, and even when all $h \in \mathcal{H}$ agree on a certain point X, $h(X)$ could be wrong. If we still run CAL algorithm, the correctness is not guaranteed.

This lecture talks about how to apply active learning in agnostic setting. We know that the basic idea of active learning is that instead of all labels, the learning algorithm only queries those it is not sure about, and infers the rest. Therefore, a specific understanding is needed to define "not sure about" in the new setting. In following paragraphs, we will firstly give an agnostic active learning algorithm, which uses error difference to measure uncertainty, then prove its correctness, and finally analyze the label complexity.

2 Algorithm

$\mathcal{X}$ is the input space and $\mathcal{Y} = \{\pm 1\}$ are the possible labels. Let s be a subset of $\mathcal{X} \times \mathcal{Y}$. Define two modified empirical risk minimization oracles:

$$\mathcal{A}(s) = \arg\min_{h \in \mathcal{H}} \mathrm{err}_s(h), \quad \mathcal{A}\left(s, \left(X', Y'\right)\right) = \arg\min_{\substack{h \in \mathcal{H} \text{ s.t.} \\ h(X') \neq Y'}} \mathrm{err}_s(h).$$

Algorithm 1 shows the agnostic active learning algorithm. Keep maintaining s_t, which is the "correct data-label pair" from learner's perspective, and apply ERM oracle $\mathcal{A}(s)$ to pick an optimal hypothesis. During each iteration, since label of X_t is either a_t or $-a_t$, the learner can check $\mathrm{err}_{s_{t-1}}(h'_{t-1})$ and $\mathrm{err}_{s_{t-1}}(h_{t-1})$ to figure out which label fits current s_{t-1} better.

As we have motioned before, the learner makes a query if and only if it is not sure about a certain example. Intuitively, a decision is hard to make from previous examples when $\mathrm{err}_{s_{t-1}}(h'_{t-1})$ and $\mathrm{err}_{s_{t-1}}(h_{t-1})$ are very close to each other. Define β_t as a given bound, "not sure about" particularly means:

$$\mathrm{err}_{s_{t-1}}(h'_{t-1}) - \mathrm{err}_{s_{t-1}}(h_{t-1}) \leq \beta_{t-1}$$

in this case.

We pick $\beta_t = \mathrm{Rad}_{t,p}(\mathcal{H}) + O\left(\sqrt{\frac{\log(1/\delta)}{t}}\right)$ mainly based on the conclusion introduced by the previous lecture on statistical learning.

Algorithm 1 Agnostic Active Learning Algorithm

Initialize $s_0 := \emptyset,\ h_0 := \mathcal{A}(s_0),\ \beta_0 := \infty$.

for $t = 1, 2, ..., n$: **do**
  Obtain unlabeled data point X_t
  Predict $a_t = h_{t-1}(X_t)$
  Obtain $h'_{t-1} = \mathcal{A}(s_{t-1}, (X_t, a_t))$
  if $\mathrm{err}_{s_{t-1}}(h'_{t-1}) \leq \mathrm{err}_{s_{t-1}}(h_{t-1}) + \beta_{t-1}$ **then**
    Get Y_t
    $s_t = s_{t-1} \cup \{(X_t, Y_t)\}$
  else
    $s_t = s_{t-1} \cup \{(X_t, a_t)\}$
  end if
  Let the bound $\beta_t = \mathrm{Rad}_{t,p}(\mathcal{H}) + O\left(\sqrt{\frac{\log(1/\delta)}{t}}\right)$, $h_t = \mathcal{A}(s_t)$
end for
return $h_n = \mathcal{A}(s_n)$.

3 Correctness analysis

Note that s_n is different from the empirical distribution on $\{(X_i, Y_i)\}_{i=1}^n$ because it has make-up data, whose label comes from inferences of the learner. Let $h^* = \arg\min_{h\in\mathcal{H}} \mathrm{err}_p(h)$ as the optimal hypothesis on the true distribution, and we need to illustrate $h_n = \mathcal{A}(s_n)$, has an error rate close to h^* in order to prove the correctness of the algorithm. Define event E_n, where p is the true unknown distribution, and p_n is the empirical distribution on $\{(X_i, Y_i)\}_{i=1}^n$:

$$\begin{cases} \max_{h\in\mathcal{H}} \left|\mathrm{err}_p(h) - \mathrm{err}_{p_n}(h)\right| \leq \mathrm{Rad}_{n,p}(\mathcal{H}) + O\left(\sqrt{\frac{\log(n(n+1)/\delta)}{n}}\right) \\ \left|\mathrm{err}_p(h^*) - \mathrm{err}_{p_n}(h^*)\right| \leq O\left(\sqrt{\frac{\log(n(n+1)/\delta)}{n}}\right) \end{cases}$$

Claim 1. *$\bigcap_{n=0}^{\infty} E_n$ hold with probability at least $1 - \delta$.*

These results come directly from previous statistical learning lecture by replacing $\frac{1}{\delta}$ with $\frac{n(n+1)}{\delta}$.

Lemma 1 (Favorable bias lemma (FBL))**.** *Suppose s_n has property: whenever a_i is used in place of Y_i, we have $a_i = h^*(X_i)$, for all $i = 1, 2, .., n$, then for any h,* $\mathrm{err}_{s_n}(h) - \mathrm{err}_{s_n}(h^*) \geq \mathrm{err}_{p_n}(h) - \mathrm{err}_{p_n}(h^*)$.

Proof. Pick i s.t. $a_i \neq Y_i$ but is used in s_n. Pick any h. If $h(X_i) = h^*(X_i)$, then $LHS = RHS = 0$. If $h(X_i) \neq h^*(X_i)$, then $h(X_i)$ at least makes one more mistake, therefore $LHS = 1 \geq RHS$. In cases when $a_i = Y_i$, obviously, $LHS = RHS$. □

Corollary 1. *Suppose s_n has the identical property in FBL, then $h_n = \arg\min_{h\in\mathcal{H}} \mathrm{err}_{s_n}(h)$ satisfies $\mathrm{err}_p(h_n) \leq \mathrm{err}_p(h^*) + \beta_n$.*

Proof.

$$\begin{aligned}\mathrm{err}_p(h_n) - \mathrm{err}_p(h^*) &\leq \mathrm{err}_{p_n}(h_n) - \mathrm{err}_{p_n}(h^*) + \beta_n \ (\textit{according to } E_n \textit{ event}) \\ &\leq \mathrm{err}_{s_n}(h_n) - \mathrm{err}_{s_n}(h^*) + \beta_n \ (\textit{according to FBL}) \\ &\leq \beta_n.\end{aligned}$$

□

By this point, we have known that the final result $\mathcal{A}(s_n)$ has an error upper bound. Let's go back to see the condition of if-statement in Algorithm 1. The lemma below gives an idea of why the learner is able to make an inference according to s_n.

Lemma 2. *Suppose $\bigcap_{n=0}^{n-1} E_n$ holds. Then*

$$\mathrm{err}_{s_m}(h_m^{'}) > \mathrm{err}_{s_m}(h_m) + \beta_m \quad \Longrightarrow \quad a_{m+1} = h^*(X_{m+1}), \qquad \forall m = 0, 1, ..., n-1 \qquad (*)$$

Proof. Use induction on m. Base case is when $m = 0$. Since $\beta_0 = \infty$, it's trivial. Assume $n \geq 1$, $\bigcap_{n=0}^{\infty} E_n$. We need to show that if $\mathrm{err}_{s_m}(h_m^{'}) > \mathrm{err}_{s_m}(h_m) + \beta_m$, then $a_{n+1} = h^*(X_{n+1})$. Consider the contrapositive: if $h^*(X_{n+1}) \neq a_{n+1}$, $a_{n+1} = h_n(X_{n+1})$, then $\mathrm{err}_{s_n}(h_n^{'}) \leq \mathrm{err}_{s_n}(h_n) + \beta_n$. First we know that

$$\mathrm{err}_{s_n}(h^*) - \mathrm{err}_{s_n}(h_n) \leq \mathrm{err}_{p_n}(h^*) - \mathrm{err}_{p_n}(h_n),$$

for (*) can reduce to FBL conditon. In addition, from event E_n,

$$\mathrm{err}_{p_n}(h^*) - \mathrm{err}_{p_n}(h_n) \leq \mathrm{err}_p(h^*) - \mathrm{err}_p(h_n) + \beta_n.$$

In all,

$$\mathrm{err}_{s_n}(h^*) - \mathrm{err}_{s_n}(h_n) \leq \mathrm{err}_p(h^*) - \mathrm{err}_p(h_n) + \beta_n \leq \beta_n.$$

By definition of $h_n^{'}$, $\mathrm{err}_{s_n}(h_n^{'}) \leq \mathrm{err}_{s_n}(h^*)$. As a result,

$$\mathrm{err}_{s_n}(h_n^{'}) \leq \mathrm{err}_{s_n}(h_n) + \beta_n.$$

□

4 Label complexity

Firstly, let's define a couple of concepts. Recall that in the realizable case, we have $\rho(h, h^{'})$, $B(h, r)$, and $\mathcal{D}(h, r)$ defined as,

$$\rho(h, h^{'}) = Pr(h(X) \neq h^{'}(X)),$$

$$B(h, r) = \left\{ h^{'} \in \mathcal{H},\ \rho(h, h^{'}) \leq r \right\},$$

$$\mathcal{D}(h,r) = \left\{ X \in \mathcal{X},\ \exists h^{'} \in B(h,r),\ h(X) \neq h^{'}(X) \right\}.$$

Similarly, in agnostic case, we need a metric to measure the difference between two hypotheses,

$$\widetilde{\rho}(h,h^{'}) = \mathrm{err}_p(h) - \mathrm{err}_p(h^{'}),$$

and also a description of a particular neighborhood,

$$\widetilde{B}(h,r) = \left\{ h \in \mathcal{H},\ \widetilde{\rho}(h,h^{'}) \leq r \right\}.$$

Introduce the concept $\widetilde{\mathcal{D}}$ to represent the examples that can distinguish h from its neighbours,

$$\widetilde{\mathcal{D}}(h,r) = \left\{ X \in \mathcal{X},\ \exists h^{'} \in \widetilde{B}(h,r),\ h^{'}(X) \neq h(X) \right\}.$$

In realizable setting, since $\mathrm{err}_p(h^*) = 0$, actually $\widetilde{\rho}(h,h^*) = \rho(h,h^*)$ always holds. In general, $\widetilde{\mathcal{D}}(h^*,r) \leq \mathcal{D}(h^*, 2\mathrm{err}_p(h^*) + r)$. This is a direct result of the fact

$$\rho(h,h^*) \leq \mathrm{err}_p(h^*) + \mathrm{err}_p(h),$$

which is supported by triangle inequality of ρ metrics.

We are already very close to revealing the label complexity of agnostic active learning algorithm. Next lemma helps us to understand when an example's label is ambiguous to the learner, and therefore, need a query.

Lemma 3. *Assume* $\bigcap_{n=0}^{n-1} E_n$ *hold, if* $\mathrm{err}_{s_{n-1}}(h^{'}_{n-1}) \leq \mathrm{err}_{s_{n-1}}(h_{n-1}) + \beta_{n-1}$, *then* $X_n \in \widetilde{\mathcal{D}}(h^*, 2\beta_{n-1}) \subseteq \mathcal{D}(h^*, 2\,\mathrm{err}_p(h^*) + 2\beta_{n-1})$.

Proof. Need to exhibit, $h \in \widetilde{B}(h^*, 2\beta_{n-1})$ s.t. $h(X_n) \neq h^*(X_n)$.

Case 1: $h_{n-1}(x_n) = h^*(X_n)$. We show that $h^{'}_{n-1}$ works, i.e. $\mathrm{err}_p(h^{'}_{n-1}) \leq \mathrm{err}_p(h^*) + 2\beta_{n-1}$:

$$\begin{aligned}
\mathrm{err}_p(h^{'}_{n-1}) - \mathrm{err}_p(h^*) &\leq \mathrm{err}_{p_{n-1}}(h^{'}_{n-1}) - \mathrm{err}_{p_{n-1}}(h^*) + \beta_{n-1} \quad \text{by } (E_{n-1}) \\
&\leq \mathrm{err}_{s_{n-1}}(h^{'}_{n-1}) - \mathrm{err}_{s_{n-1}}(h^*) + \beta_{n-1} \quad \text{by (FBL)} \\
&\leq \mathrm{err}_{s_{n-1}}(h^{'}_{n-1}) - \mathrm{err}_{s_{n-1}}(h_{n-1}) + \beta_{n-1} \\
&\leq 2\beta_{n-1}
\end{aligned}$$

Case 2: $h_{n-1}(x_n) \neq h^*(X_n)$. Basically the same as Case 1 for h_{n-1} is ERM. □

This lemma can be used in exactly the same fashion as in the realizable setting for bounding the expected number of labels requested.

Bibliographic notes

The agnostic active learning algorithm is due to Sanjoy Dasgupta, Daniel Hsu, and Claire Monteleoni [1], while the material is quite different from the lecture given by Prof. Hsu. This scibe is mainly written according to Prof. Hsu's lecture. Event E_n comes from statistical learning theory material contributed by Bousquet, Boucheron, and Lugosi [2]. Basic realizable setting active learning algorithm and CAL algorithm is introduced in Cohn, Atlas, and Ladner [3]'s paper

References

[1] Dasgupta, Hsu, and Monteleoni. A general agnostic active learning algorithm. *Advances in neural information processing systems*, 353-360, 2008.

[2] Bousquet, Boucheron, and Lugosi. Introduction to Statistical Learning Theory. *Springer Berlin Heidelberg*, 169-207, 2004.

[3] Cohn, Atlas, and Ladner. Improving generalization with active learning. *Machine Learning*, 15: 201, 1994.