Agnostic Active Learning Notes

Kevin Jamieson

November 26, 2010

Agnostic learning is one of those most general learning settings in which little or no assumptions are made about the true underlying function. Some even extend this definition to adversarial noise but we assume we are streamed an iid sample from some unknown joint distribution over $X \times \{-1, 1\}$.

The algorithm and proofs closely follow Dasgupta, Hsu, and Monteleoni's paper: A General Agnostic Active Learning Algorithm.

Definition 1. *For a hypothesis class $\mathcal{H}$ and dataset $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in X$ and $y_i \in \{-1, 1\}$ we call $\hat{h}$ the empirical risk minimizer if $\hat{h} = \arg\min_{h \in \mathcal{H}} \mathrm{err}_{\mathrm{n}}(h)$ where*

$$\mathrm{err}_{\mathrm{n}}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{h(x_i) \neq y_i\}$$

Definition 2. *If a finite dataset is identically and independently distributed from some distribution, that is, $\{(x_i, y_i)\}_{i=1}^n \sim \mathcal{D}_{X,Y}$ then the true risk of a hypothesis $h \in \mathcal{H}$ is denoted*

$$\mathrm{err}_{\mathcal{D}}(h) = P(h(X) \neq Y) = \int_X \mathbf{1}\{h(X) \neq Y\} dP(X)$$

Definition 3. *After observing n examples we call the set of all hypotheses still "in the running" to be the true risk minimizer the version space and denote it $V_n \subset \mathcal{H}$.*

We define the version space for the agnostic setting to be those hypotheses that we cannot reject as being the minimum risk classifier with high probability. More formally, if $f = \arg\inf_{h \in \mathcal{H}} \mathrm{err}_{\mathcal{D}}(h)$ and $\mathrm{err}_{\mathcal{D}}(f) = \nu$ then for some confidence $\delta > 0$:

$$V_n = \{h \in \mathcal{H} : P(\mathrm{err}_{\mathcal{D}}(h) > \nu) \leq \delta\}$$

for all $n \in \mathbb{N}$. It follows that with probability $> 1 - \delta$, $f \in V_n$.

Theorem 1. *Vapnik (1971) For a hypothesis class $\mathcal{H}$ with finite VC dimension d, $\alpha_n = \sqrt{\frac{4d \log 2n + 4 \log(8/\delta)}{n}}$, and an iid sample $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{D}_{X,Y}$ then for all $h \in \mathcal{H}$ with probability greater than $1 - \delta$*

$$-\min\left\{\alpha_n \sqrt{\mathrm{err}_{\mathrm{n}}(h)}, \alpha_n^2 + \alpha_n \sqrt{\mathrm{err}_{\mathcal{D}}(h)}\right\} \leq \mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_{\mathrm{n}}(h) \leq \min\left\{\alpha_n^2 + \alpha_n \sqrt{\mathrm{err}_{\mathrm{n}}(h)}, \alpha_n \sqrt{\mathrm{err}_{\mathcal{D}}(h)}\right\}.$$

Note that the above theorem can be more useful than a bound on just $|\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_{\mathrm{n}}(h)|$.

Lemma 1. *Dasgupta (2007) For a hypothesis class $\mathcal{H}$ with finite VC dimension d, $\beta_n = \sqrt{\frac{8d \log 2n + 4 \log(8(n^2+n)/\delta)}{n}}$, and an iid sample $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{D}_{X,Y}$ then for all $h, h' \in \mathcal{H}$ with probability greater than $1 - \delta$*

$$\mathrm{err}_{\mathrm{n}}(h) - \mathrm{err}_{\mathrm{n}}(h') \leq \mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_{\mathcal{D}}(h') + \beta_n^2 + \beta_n(\sqrt{\mathrm{err}_{\mathrm{n}}(h)} + \sqrt{\mathrm{err}_{\mathrm{n}}(h')}).$$

If for some $h, h' \in \mathcal{H}$, $\mathrm{err}_{\mathrm{n}}(h) - \mathrm{err}_{\mathrm{n}}(h') > \Delta_n$ then $\mathrm{err}_{\mathcal{D}}(h) > \mathrm{err}_{\mathcal{D}}(h')$ where

$$\Delta_n = \beta_n^2 + \beta_n(\sqrt{\mathrm{err}_{\mathrm{n}}(h)} + \sqrt{\mathrm{err}_{\mathrm{n}}(h')}). \tag{1}$$

This inequality is novel in itself and may be very useful in maintaining a version space.

Definition 4. *For some hypothesis class $\mathcal{H}$ and set $\mathcal{X}$ where for $h \in \mathcal{H}$, $h : \mathcal{X} \to \{-1, 1\}$, the region of disagreement is defined as*

$$DIS(\mathcal{H}) = \{x \in \mathcal{X} : \exists h, h' \in \mathcal{H} \text{ s.t. } h(x) \neq h'(x)\}.$$

Definition 5. *For some hypothesis class $\mathcal{H}$ and the marginal $\mathcal{D}_X$ of $\mathcal{D}_{X,Y}$ the closed ball centered at $h \in \mathcal{H}$ with radius r is defined as*

$$B(h, r) = \{h' \in \mathcal{H} : \rho(h, h') \leq r\}.$$

where $\rho(h, h') = P_{X\sim\mathcal{D}}(h(X) \neq h'(X))$.

Definition 6. *The disagreement coefficient $\theta_f = \theta(\mathcal{H}, \mathcal{D}_{X,Y}, \epsilon)$ of $f \in \mathcal{H}$ with respect to $\mathcal{H}$ and $\mathcal{D}_{X,Y}$ is*

$$\theta_f = \sup_{\epsilon>0} \frac{P(DIS(B(f, \nu + \epsilon)))}{\nu + \epsilon}.$$

Algorithm 1.
Input: stream $(x_1, x_2, \ldots, x_m)$ iid from $\mathcal{D}_X$
Initially, $\hat{S}_0 = \emptyset$ and $T_0 = \emptyset$
For $n = 1, 2, \ldots, m$:

1. For each $\hat{y} \in \{\pm 1\}$, let $h_{\hat{y}} = \text{LEARN}_{\mathcal{H}}(\hat{S}_{n-1} \cup \{(x_n, \hat{y})\}, T_{n-1})$.
2. If $\text{err}_{\text{n}}(h_{-\hat{y}}) - \text{err}_{\text{n}}(h_{\hat{y}}) > \Delta_{n-1}$ (or if no such $h_{-\hat{y}}$ is found) for some $\hat{y} \in \{\pm 1\}$, then $\hat{S}_n = \hat{S}_{n-1} \cup \{(x_n, \hat{y})\}$ and $T_n = T_{n-1}$.
3. Else request y_n; $\hat{S}_n = \hat{S}_{n-1}$ and $T_n = T_{n-1} \cup \{(x_n, y_n)\}$.

where $\text{LEARN}_{\mathcal{H}}(\hat{S}_{n-1} \cup \{(x_n, \hat{y})\}, T_{n-1})$ is a subroutine that learns a classifier that is consistent with its first argument and achieves minimum risk on its second argument.

Observation 1. In the above $\text{LEARN}(\hat{S}, T)$ subroutine, $\hat{S}$ include imputed labels determined by the algorithm that need not equal the true labels in S. If $\widehat{\text{err}}_{\text{n}}\, h$ denotes the empirical error on $\hat{S} \cup T$ and $\text{err}_{\text{n}}(h)$ is the empirical error on $S \cup T$ then for any two hypotheses $h, h' \in V_n$

$$\widehat{\text{err}}_{\text{n}}(h) - \widehat{\text{err}}_{\text{n}}(h') = \text{err}_{\text{n}}(h) - \text{err}_{\text{n}}(h'). \tag{2}$$

It then follows that we can replace all $\text{err}_{\text{n}}(h)$ terms with $\widehat{\text{err}}_{\text{n}}(h)$ terms in Lemma 1.

Lemma 2. *If Algorithm 1 is streamed an iid sample $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{D}_{X,Y}$ and $f = \arg\inf_{h\in\mathcal{H}} \text{err}_{\mathcal{D}}(h)$ then with probability greater than $1 - \delta$ $f \in V_n$ and f is consistent with $\hat{S}_n$.*

Proof. Suppose we are given some x_n and we find $\widehat{\text{err}}_{\text{n}}(h_{+1}) - \widehat{\text{err}}_{\text{n}}(h_{-1}) > \Delta_n$. We would then add $(x_n, -1)$ to $\hat{S}_n$ without requesting a label and move on. We will now show $f(x_n) = -1$.

Suppose not. Then $f(x_n) = +1$ and because h_{+1} is the empirical risk minimizer over $\hat{S}_n \cup T_n$ by construction, $\widehat{\text{err}}_{\text{n}}(f) \geq \widehat{\text{err}}_{\text{n}}(h_{+1})$. Recalling that $\widehat{\text{err}}_{\text{n}}(h_{+1}) - \widehat{\text{err}}_{\text{n}}(h_{-1}) > \beta_n^2 + \beta_n(\sqrt{\widehat{\text{err}}_{\text{n}}(h)} + \sqrt{\widehat{\text{err}}_{\text{n}}(h')})$ we see

$$\begin{aligned}
\widehat{\text{err}}_{\text{n}}(f) - \widehat{\text{err}}_{\text{n}}(h_{-1}) &= \widehat{\text{err}}_{\text{n}}(f) - \widehat{\text{err}}_{\text{n}}(h_{+1}) + \widehat{\text{err}}_{\text{n}}(h_{+1}) - \widehat{\text{err}}_{\text{n}}(h_{-1}) \\
&> \sqrt{\widehat{\text{err}}_{\text{n}}(h_{+1})}[\sqrt{\widehat{\text{err}}_{\text{n}}(f)} - \sqrt{\widehat{\text{err}}_{\text{n}}(h_{+1})}] + \Delta_n \\
&> \beta_n[\sqrt{\widehat{\text{err}}_{\text{n}}(f)} - \sqrt{\widehat{\text{err}}_{\text{n}}(h_{+1})}] + \beta_n^2 + \beta_n[\sqrt{\widehat{\text{err}}_{\text{n}}(h_{+1})} - \sqrt{\widehat{\text{err}}_{\text{n}}(h_{-1})}] \\
&= \beta_n^2 + \beta_n[\sqrt{\widehat{\text{err}}_{\text{n}}(h_f)} + \sqrt{\widehat{\text{err}}_{\text{n}}(h_{-1})}]
\end{aligned}$$

but this implies $\text{err}_{\mathcal{D}}(f) > \text{err}_{\mathcal{D}}(h_{-1})$ by Lemma 1, which is a contradiction. □

Theorem 2. *If Algorithm 1 is streamed an iid sample $\{(x_i, y_i)\}_{i=1}^n$ from $\mathcal{D}_{X,Y}$, $f = \arg\inf_{h\in\mathcal{H}} \text{err}_{\mathcal{D}}(h)$ and $\text{err}_{\mathcal{D}}(f) = \nu$, and $\hat{h} = \arg\min_{h\in\mathcal{H}} \widehat{\text{err}}_{\text{n}}(h)$ then with probability greater than $1 - \delta$*

$$\text{err}_{\mathcal{D}}(\hat{h}) \leq \nu + \frac{24d\log 2n + 12\log(8(n^2+n)/\delta)}{n} + \sqrt{\nu}\sqrt{\frac{32d\log 2n + 16\log(8(n^2+n)/\delta)}{n}}.$$

Proof. We use Theorem 1, the fact that $\widehat{\text{err}}_{\text{n}}(\hat{h}) \leq \widehat{\text{err}}_{\text{n}}(f)$, and $\alpha_n \leq \beta_n$ to see

$$\begin{aligned}
\text{err}_{\mathcal{D}}(\hat{h}) &\leq \widehat{\text{err}}_{\text{n}}(\hat{h}) - \widehat{\text{err}}_{\text{n}}(f) + \nu + \beta_n^2 + \beta_n\sqrt{\text{err}_{\mathcal{D}}(\hat{h})} + \beta_n\sqrt{\nu} \\
&\leq \nu + \beta_n^2 + \beta_n\sqrt{\text{err}_{\mathcal{D}}(\hat{h})} + \beta_n\sqrt{\nu} \\
&\leq \nu + 3\beta_n^2 + 2\beta_n\sqrt{\nu}
\end{aligned}$$

where the last line comes from the inequality:

$$A \leq B + C\sqrt{A} \implies A \leq B + C^2 + C\sqrt{B}$$

for non-negative A, B, C. □

In the realizable setting our strategy was to find the number of labels required to achieve an ϵ excess risk. In this agnostic case, we determine the number of unlabeled examples necessary to achieve a certain ϵ and then determine the expected number of labels requested.

Suppose $f = \arg\inf_{h\in\mathcal{H}} \text{err}_{\mathcal{D}}(h)$ and $f(x_n) = -1$. Then a label is requested if $\widehat{\text{err}}_{\text{n}}(h_{+1}) - \widehat{\text{err}}_{\text{n}}(h_{-1}) > \beta_n^2 + \beta_n(\sqrt{\widehat{\text{err}}_{\text{n}}(h)} + \sqrt{\widehat{\text{err}}_{\text{n}}(h')})$. How small does $\text{err}_{\mathcal{D}}(h_{+1})$ have to be to cause this event to occur and how likely is that to happen as n gets big? How does this relate to the disagreement coefficient? This leads us to our next two lemmas.

Lemma 3. *Let $f = \arg\inf_{h\in\mathcal{H}} \text{err}_{\mathcal{D}}(h)$, $\text{err}_{\mathcal{D}}(f) = \nu$, $f(x_{n+1}) = \hat{y}$, and y_{n+1} is requested. Then with probability $\geq 1 - 2\delta$*

$$\text{err}_{\mathcal{D}}(h_{-\hat{y}}) \leq 3\nu + (12 + 2\sqrt{3})\beta_n^2$$

and for some $c_1, c_2 > 0$

$$P(\textit{Request } y_{n+1}) \leq P(DIS(B(f, (c_1+1)\nu + c_2\beta_n^2)))$$

Proof. The first part is proved by pushing inequalities around and isn't very instructive. For the second part, define $\tilde{h}$ to be the Bayes decision boundary so that with metric ρ from Definition 5 we see that for any $h \in \mathcal{H}$ $\rho(\tilde{h}, h) \leq \text{err}_{\mathcal{D}}(h) - \text{err}_{\mathcal{D}}(\tilde{h}) \leq \text{err}_{\mathcal{D}}(h)$. It then follows that

$$\begin{aligned}
\rho(h_{-\hat{y}}, f) &\leq \rho(h_{-\hat{y}}, \tilde{h}) + \rho(\tilde{h}, f) \\
&\leq \text{err}_{\mathcal{D}}(h_{-\hat{y}}) + \text{err}_{\mathcal{D}}(f) \\
&= \text{err}_{\mathcal{D}}(h_{-\hat{y}}) + \nu.
\end{aligned}$$

□

Lemma 4. *With the same conditions of Lemma 3, there exists a constant $c > 0$ such that*

$$P(\textit{Request } y_{n+1}) \leq c \cdot \theta_f(\nu + \beta_n^2)$$

where $\theta_f = \theta(\mathcal{D}, \mathcal{H}, \epsilon_n)$ and $\epsilon_n = 3\beta_n^2 + 2\beta_n\sqrt{\nu}$, the achievable excess risk with n examples.

Proof. We can choose constants $c_1, c_2 > 0$ such that $c_1\nu + c_2\beta_n^2 \geq 3\beta_n^2 + 2\beta_n\sqrt{\nu}$. We then find

$$\begin{aligned}
\frac{P(\text{Request } y_{n+1})}{(c_1+1)\nu + c_2\beta_n^2} &\leq \frac{P(DIS(B(f, (c_1+1)\nu + c_2\beta_n^2)))}{(c_1+1)\nu + c_2\beta_n^2} \\
&\leq \sup_{\epsilon \geq \epsilon_n} \frac{P(DIS(B(f, \nu+\epsilon)))}{\nu + \epsilon} \\
&= \theta_f
\end{aligned}$$

which implies

$$\begin{aligned}
P(\text{Request } y_{n+1}) &\leq \theta_f \cdot ((c_1+1)\nu + c_2\beta_n^2) \\
&\leq c \cdot \theta_f(\nu + \beta_n^2)
\end{aligned}$$

for some $c > 0$. □

Theorem 3. *If Algorithm 1 is provided $n = \tilde{O}(\theta_f d \frac{\nu+\epsilon}{\epsilon^2})$[1] unlabeled examples from $\mathcal{D}$ then with probability $\geq 1-\delta$ a classifier is returned with risk at most $\nu + \epsilon$. Furthermore, if $\mathcal{L}_\epsilon$ denotes the expected number of labels requested*

$$\mathcal{L}_\epsilon = \begin{cases} \tilde{O}\big(\theta_f d \log^2(1/\epsilon)\big), & \text{if } \epsilon \approx \nu \\ \tilde{O}\big(\theta_f d\,[\log^2(1/\epsilon) + (\nu/\epsilon)^2]\big), & \text{if } \epsilon \ll \nu \end{cases}$$

Proof. The number of unlabeled data is found by solving for n from the risk bound in Theorem 2. To bound the number of expected labels:

$$\begin{aligned} \mathcal{L}_\epsilon &= E\Big[\sum_{t=1}^{n} \mathbf{1}\{Request\, y_t\}\Big] \\ &= \sum_{t=1}^{n} P(Request\, y_t) \\ &\leq \sum_{t=1}^{n} c \cdot \theta_f(\nu + \beta_t^2). \end{aligned}$$

Suppose $\nu \leq \epsilon_n = 3\beta_n^2 + 2\beta_n\sqrt{\nu}$ then by the inequality in the proof of Theorem 2, $\nu \leq (7+2\sqrt{3})\beta_n^2$ and

$$\begin{aligned} \mathcal{L}_\epsilon &\leq \sum_{t=1}^{n} c \cdot \theta_f\big((7+2\sqrt{3})\beta_n^2 + \beta_t^2\big) \\ &\leq c \cdot \theta_f\big((7+2\sqrt{3})\beta_n^2 n + \sum_{t=1}^{n} \beta_t^2\big) \\ &= O\big(\theta_f\big(d\log^2 n + \log(1/\delta)\log n\big)\big) \\ &= O\big(\theta_f d \log^2(1/\epsilon) + \theta_f \log(1/\delta)\log(1/\epsilon)\big). \end{aligned}$$

The other proof is straightforward. □

Things to consider

- We know $P(Request\, y_t) \leq c \cdot \theta_f(\nu + \beta_t^2)$ for $c > 1$ but what if $\theta_f \cdot \nu > 1$?
- What is the computational complexity of Algorithm 1?
- Known lower bounds?

[1] $\tilde{O}(f)$ ignores any constants, $\log(1/\delta)$, or $\log(\log(1/\epsilon))$ factors