Efficient PAC Learning from the Crowd

Pranjal Awasthi,Avrim Blum,Nika Haghtalab,Yishay Mansour

COLT-2017

Background

A crowdsourced model uses a large pool of workers to gather labels for a given training data set that will be used for the purpose of learning a good classifier. Such learning environments that involve the crowd give rise to a multitude of design choices that do not appear in traditional learning environments.

How to efficiently learn and generalize from the crowd with minimal cost?

Background

定义 12.1 PAC 辨识 (PAC Identify): 对 $0 < \epsilon, \delta < 1$, 所有 $c \in \mathcal{C}$ 和分布 $\mathcal{D}$, 若存在学习算法 $\mathfrak{L}$, 其输出假设 $h \in \mathcal{H}$ 满足

$$P(E(h) \leqslant \epsilon) \geqslant 1 - \delta \ , \tag{12.9}$$

则称学习算法 $\mathfrak{L}$ 能从假设空间 $\mathcal{H}$ 中 PAC 辨识概念类 $\mathcal{C}$.

定义 12.2 PAC 可学习 (PAC Learnable): 令 m 表示从分布 $\mathcal{D}$ 中独立同分布采样得到的样例数目, $0 < \epsilon, \delta < 1$, 对所有分布 $\mathcal{D}$, 若存在学习算法 $\mathfrak{L}$ 和多项式函数 $\text{poly}(\cdot, \cdot, \cdot, \cdot)$, 使得对于任何 $m \geqslant \text{poly}(1/\epsilon, 1/\delta, \text{size}(\boldsymbol{x}), \text{size}(c))$, $\mathfrak{L}$ 能从假设空间 $\mathcal{H}$ 中 PAC 辨识概念类 $\mathcal{C}$, 则称概念类 $\mathcal{C}$ 对假设空间 $\mathcal{H}$ 而言是 PAC 可学习的, 有时也简称概念类 $\mathcal{C}$ 是 PAC 可学习的.

Notations

- X Instance space
- $Y = \{+1, -1\}$ Label space
- $f: X \to Y$ Hypothesis
- $f^*: err_D(f^*) = 0$ Target function in hypothesis class F
- D Distribution over $X \times Y$
- $D_{|X}$ The marginal of D over X
- $err_D(f) = \Pr_{(x, f^*(x)) \sim D}[f(x) \neq f^*(x)]$

Notations

- $g_i : X \to Y$ Define labeler i, we say that g_i is perfect if $err_D(g_i) = 0$
- P The distribution over all labelers which is uniform
- $\alpha = \Pr_{i \sim P}[err_D(g_i) = 0]$ The fraction of perfect labelers
- $m_{\varepsilon,\delta}$ The total number of labeled samples drawn from the realizable distribution D needed to output a classifier f that has $errD(f) \leq \varepsilon$, with probability 1-δ. We know from the VC theory, that for a hypothesis class F with VC-dimension d and no additional assumptions on F, $m_{\varepsilon,\delta} \in O(\varepsilon^{-1}(d \ln(1/\varepsilon) + \ln(1/\delta)))$

Notations

- L A set of labelers
- $Maj_L(x)$ The label assigned to x by the majority of labelers in L
- $MAJ(L)$ The classifier that for each x returns prediction $Maj_L(x)$

When $\alpha = \frac{1}{2} + \Theta(1)$, per sample needs $O(\log(\frac{m}{\delta}))$ queries to obtain almost perfect label where m is the training data size and δ is the desired failure probability

Algorithm

Consider a very simple baseline algorithm for the case of $\alpha > 0.5$.

BASELINE: Draw a sample of size $m = m_{\epsilon,\delta}$ from $D_{|\mathcal{X}}$ and label each example x by $\mathrm{Maj}_L(x)$, where $L \sim P^k$ for $k = O\left((\alpha - 0.5)^{-2} \ln\left(\frac{m}{\delta}\right)\right)$ is a set of randomly drawn labelers. Let S be the resulting labeled set. Return classifier $\mathcal{O}_{\mathcal{F}}(S)$.

1. take $\log(m/\delta)$ more labels than it requires samples

2. when perfect labelers form a small majority of the labelers

Algorithm

Boosting

Theorem 1 (Schapire (1990)) *For any $p < \frac{1}{2}$ and distribution D, consider three classifiers: 1) classifier h_1 such that $\mathrm{err}_D(h_1) \leq p$; 2) classifier h_2 such that $\mathrm{err}_{D_2}(h_2) \leq p$, where $D_2 = \frac{1}{2}D_C + \frac{1}{2}D_I$ for distributions D_C and D_I that denote distribution D conditioned on $\{x \mid h_1(x) = f^*(x)\}$ and $\{x \mid h_1(x) \neq f^*(x)\}$, respectively; 3) classifier h_3 such that $\mathrm{err}_{D_3}(h_3) \leq p$, where D_3 is D conditioned on $\{x \mid h_1(x) \neq h_2(x)\}$. Then, $\mathrm{err}_D(MAJ(h_1, h_2, h_3)) \leq 3p^2 - 2p^3$.*

$$h_1 \sim D_1, h_2 \sim D_2 = \frac{1}{2}D_I + \frac{1}{2}D_C, h_3 \sim D_3$$

$$err_{D_i}(h_i) \leq p, i = 1,2,3 \Rightarrow err_D(MAJ(h_1, h_2, h_3)) \leq 3p^2 - 2p^3$$

Algorithm

Algorithm 2 INTERLEAVING: BOOSTING BY PROBABILISTIC FILTERING FOR $\alpha = \frac{1}{2} + \Theta(1)$

Input: Given a distribution $D_{|\mathcal{X}}$, a class of hypotheses $\mathcal{F}$, parameters ϵ and δ.

Phase 1:

Let $\overline{S_1} = \text{CORRECT-LABEL}(S_1, \delta/6)$, for a set of sample S_1 of size $2m_{\sqrt{\epsilon},\delta/6}$ from $D_{|\mathcal{X}}$.

Let $h_1 = \mathcal{O}_{\mathcal{F}}(\overline{S_1})$.

Phase 2:

Let $S_I = \text{FILTER}(S_2, h_1)$, for a set of samples S_2 of size $\Theta(m_{\epsilon,\delta})$ drawn from $D_{|\mathcal{X}}$.

Let S_C be a sample set of size $\Theta(m_{\sqrt{\epsilon},\delta})$ drawn from $D_{|\mathcal{X}}$.

Let $\overline{S_{All}} = \text{CORRECT-LABEL}(S_I \cup S_C, \delta/6)$.

Let $\overline{W_I} = \{(x, y) \in \overline{S_{All}} \mid y \neq h_1(x)\}$ and Let $\overline{W_C} = \overline{S_{All}} \setminus \overline{W_I}$.

Draw a sample set $\overline{W}$ of size $\Theta(m_{\sqrt{\epsilon},\delta})$ from a distribution that equally weights $\overline{W_I}$ and $\overline{W_C}$.

Let $h_2 = \mathcal{O}_{\mathcal{F}}(\overline{W})$.

Phase 3:

Let $\overline{S_3} = \text{CORRECT-LABEL}(S_3, \delta/6)$, for a sample set S_3 of size $2m_{\sqrt{\epsilon},\delta/6}$ drawn from $D_{|\mathcal{X}}$ conditioned on $h_1(x) \neq h_2(x)$.

Let $h_3 = \mathcal{O}_{\mathcal{F}}(\overline{S_3})$.

return $\text{Maj}(h_1, h_2, h_3)$.

Algorithm

Input: Given a distribution $D_{|\mathcal{X}}$, a class of hypotheses $\mathcal{F}$, parameters ϵ and δ.

Phase 1:

Let $\overline{S_1} = \textsc{Correct-Label}(S_1, \delta/6)$, for a set of sample S_1 of size $2m_{\sqrt{\epsilon},\delta/6}$ from $D_{|\mathcal{X}}$.

Let $h_1 = \mathcal{O}_{\mathcal{F}}(\overline{S_1})$.

$\textsc{Correct-Label}(S, \delta)$:

---

for $x \in S$ **do**

Let $L \sim P^k$ for a set of $k = O(\log(\frac{|S|}{\delta}))$ labelers drawn from P and $\overline{S} \leftarrow \overline{S} \cup \{(x, \mathrm{Maj}_L(x))\}$.

end

return $\overline{S}$.

$$m_{c\varepsilon,\delta} = O(\frac{1}{c} m_{\varepsilon,\delta}), \Pr(err_D(h_1) \leq \frac{1}{2}\sqrt{\varepsilon}) > 1 - \frac{\delta}{3}$$

Algorithm

Phase 2:

Let $S_I = \text{FILTER}(S_2, h_1)$, for a set of samples S_2 of size $\Theta(m_{\epsilon,\delta})$ drawn from $D_{|\mathcal{X}}$.
Let S_C be a sample set of size $\Theta(m_{\sqrt{\epsilon},\delta})$ drawn from $D_{|\mathcal{X}}$.
Let $\overline{S_{All}} = \text{CORRECT-LABEL}(S_I \cup S_C, \delta/6)$.
Let $\overline{W_I} = \{(x, y) \in \overline{S_{All}} \mid y \neq h_1(x)\}$ and Let $\overline{W_C} = \overline{S_{All}} \setminus \overline{W_I}$.
Draw a sample set $\overline{W}$ of size $\Theta(m_{\sqrt{\epsilon},\delta})$ from a distribution that equally weights $\overline{W_I}$ and $\overline{W_C}$.
Let $h_2 = \mathcal{O}_{\mathcal{F}}(\overline{W})$.

$$\overline{W} \neq D_2$$

Algorithm

Algorithm 1 FILTER(S, h)

Let $S_I = \emptyset$ and $N = \log\left(\frac{1}{\epsilon}\right)$.
for $x \in S$ **do**
    for $t = 1, \ldots, N$ **do**
        Draw a random labeler $i \sim P$ and let $y_t = g_i(x)$
        If t is odd and $\mathrm{Maj}(y_{1:t}) = h(x)$, **then** break.
    end
    Let $S_I = S_I \cup \{x\}$. // Reaches this step when for all t, $\mathrm{Maj}(y_{1:t}) \neq h(x)$
end
return S_I

Algorithm

Lemma 6 *Given any sample set S and classifier h, for every $x \in S$*

1. *If $h(x) = f^*(x)$, then $x \in \text{FILTER}(S, h)$ with probability $< \sqrt{\epsilon}$.*
2. *If $h(x) \neq f^*(x)$, then $x \in \text{FILTER}(S, h)$ with probability ≥ 0.5.*

Proof For the first claim, note that $x \in S_I$ only if $\text{Maj}(y_{1:t}) \neq h(x)$ for all $t \leq N$. Consider $t = N$ time step. Since each random query agrees with $f^*(x) = h(x)$ with probability ≥ 0.7 independently, majority of $N = O(\log(1/\sqrt{\epsilon}))$ labels are correct with probability at least $1 - \sqrt{\epsilon}$. Therefore, the probability that the majority label disagrees with $h(x) = f^*(x)$ at every time step is at most $\sqrt{\epsilon}$.

Algorithm

2. *If* $h(x) \neq f^*(x)$, *then* $x \in \text{FILTER}(S, h)$ *with probability* ≥ 0.5.

In the second claim, we are interested in the probability that there exists some $t \leq N$, for which $\text{Maj}(y_{1:t}) = h(x) \neq f^*(x)$. This is the same as the probability of return in biased random walks, also called the probability of ruin in gambling (Feller, 2008), where we are given a random walk that takes a step to the right with probability ≥ 0.7 and takes a step to the left with the remaining probability and we are interested in the probability that this walk ever crosses the origin to the left while taking N or even infinitely many steps. Using the probability of return for biased random walks (see Theorem 15), the probability that $\text{Maj}(y_{1:t}) \neq f^*(x)$ ever is at most $\left(1 - \left(\frac{0.7}{1-0.7}\right)^N\right) / \left(1 - \left(\frac{0.7}{1-0.7}\right)^{N+1}\right) < \frac{3}{7}$. Therefore, for each x such that $h(x) \neq f^*(x)$, $x \in S_I$ with probability at least $4/7$. ■

Algorithm

Lemma 7 *With probability* $1 - \exp(-\Omega(m_{\sqrt{\epsilon},\delta}))$, $\overline{W}_I$, $\overline{W}_C$, *and* S_I *all have size* $\Theta(m_{\sqrt{\epsilon},\delta})$.

Let us first consider the expected size of sets S_I, $\overline{W}_I$, and $\overline{W}_C$. Using Lemma 6, we have

$$O(m_{\sqrt{\epsilon},\delta}) \geq \frac{1}{2}\sqrt{\epsilon}|S_2| + \sqrt{\epsilon}|S_2| \geq \mathbb{E}[|S_I|] \geq \frac{1}{2}\left(\frac{1}{2}\sqrt{\epsilon}\right)|S_2| \geq \Omega(m_{\sqrt{\epsilon},\delta}).$$

Similarly,

$$O(m_{\sqrt{\epsilon},\delta}) \geq \mathbb{E}[S_I] + |S_C| \geq \mathbb{E}[\overline{W}_I] \geq \frac{1}{2}\left(\frac{1}{2}\sqrt{\epsilon}\right)|S_2| \geq \Omega(m_{\sqrt{\epsilon},\delta}).$$

Similarly,

$$O(m_{\sqrt{\epsilon},\delta}) \geq \mathbb{E}[S_I] + |S_C| \geq \mathbb{E}[\overline{W}_C] \geq \left(1 - \frac{1}{2}\sqrt{\epsilon}\right)|S_C| \geq \Omega(m_{\sqrt{\epsilon},\delta}).$$

The claim follows by the Chernoff bound.

Algorithm

Lemma 2 *Given a hypothesis class $\mathcal{F}$ consider any two discrete distributions D and D' such that for all x, $\rho'(x) \geq c \cdot \rho(x)$ for an absolute constant $c > 0$, and both distributions are labeled according to $f^* \in \mathcal{F}$. There exists a constant $c' > 1$ such that for any ϵ and δ, with probability $1 - \delta$ over a labeled sample set S of size $c' m_{\epsilon,\delta}$ drawn from D', $\mathcal{O}_{\mathcal{F}}(S)$ has error of at most ϵ with respect to distribution D.*

Algorithm

First, notice that because D and D' are both labeled according to $f^* \in \mathcal{F}$, for any $f \in \mathcal{F}$ we have,

$$\mathrm{err}_{D'}(f) = \sum_x \rho'(x)\mathbb{1}_{f(x)\neq f^*(x)} \geq \sum_x c \cdot \rho(x)\mathbb{1}_{f(x)\neq f^*(x)} = c \cdot \mathrm{err}_D(f).$$

Therefore, if $\mathrm{err}_{D'}(f) \leq c\epsilon$, then $\mathrm{err}_D(f) \leq \epsilon$. Let $m' = m_{c\epsilon,\delta}$, we have

$$\begin{aligned}\delta &> \Pr_{S'\sim D'^{m'}}[\exists f \in \mathcal{F}, \text{s.t. } \mathrm{err}_{S'}(f) = 0 \wedge \mathrm{err}_{D'}(f) \geq c\epsilon] \\ &\geq \Pr_{S'\sim D'^{m'}}[\exists f \in \mathcal{F}, \text{s.t. } \mathrm{err}_{S'}(f) = 0 \wedge \mathrm{err}_{D}(f) \geq \epsilon].\end{aligned}$$

The claim follows by the fact that $m_{c\epsilon,\delta} = O\left(\frac{1}{c}m_{\epsilon,\delta}\right)$.

Algorithm

Lemma 8 *Let D_C and D_I denote distribution D when it is conditioned on $\{x \mid h_1(x) = f^*(x)\}$ and $\{x \mid h_1(x) \neq f^*(x)\}$, respectively, and let $D_2 = \frac{1}{2}D_I + \frac{1}{2}D_C$. With probability $1 - 2\delta/3$, $\mathrm{err}_{D_2}(h_2) \leq \frac{1}{2}\sqrt{\epsilon}$.*

Let $\rho(x)$, $\rho_2(x)$, $\rho_C(x)$, and $\rho_I(x)$ be the density of instance x in distributions D, D_2, D_C *and* D_I , respectively.

Let $N_C(x)$, $N_I(x)$, $M_C(x)$ and $M_I(x)$ be the number of occurrences of x in the sets S_C, S_I, $\overline{W_C}$ and $\overline{W_I}$, respectively.

Algorithm

If $h_1(x) = f^*(x)$: Then, there exist absolute constants c_1 and c_2 according to Lemma 7, such that

$$\rho'(x) = \frac{1}{2}\mathbb{E}\left[\frac{M_C(x)}{|\overline{W_C}|}\right] \geq \frac{\mathbb{E}[M_C(x)]}{c_1 \cdot m_{\sqrt{\epsilon},\delta}} \geq \frac{\mathbb{E}[N_C(x)]}{c_1 \cdot m_{\sqrt{\epsilon},\delta}} = \frac{|S_C| \cdot \rho(x)}{c_1 \cdot m_{\sqrt{\epsilon},\delta}}$$

$$= \frac{|S_C| \cdot \rho_C(x) \cdot (1 - \frac{1}{2}\sqrt{\epsilon})}{c_1 \cdot m_{\sqrt{\epsilon},\delta}} \geq c_2\rho_C(x) = \frac{c_2\rho_2(x)}{2},$$

$$\rho(x) = \rho_C(x)(1 - \tfrac{1}{2}\sqrt{\epsilon})$$

If $h_1(x) \neq f^*(x)$: Then, there exist absolute constants c_1' and c_2' according to Lemma 7, such that

$$\rho'(x) = \frac{1}{2}\mathbb{E}\left[\frac{M_I(x)}{|\overline{W_I}|}\right] \geq \frac{\mathbb{E}[M_I(x)]}{c_1' \cdot m_{\sqrt{\epsilon},\delta}} \geq \frac{\mathbb{E}[N_I(x)]}{c_1' \cdot m_{\sqrt{\epsilon},\delta}} \geq \frac{\frac{4}{7}\,\rho(x)|S_2|}{c_1' \cdot m_{\sqrt{\epsilon},\delta}}$$

$$= \frac{\frac{4}{7}\,\rho_I(x)\frac{1}{2}\sqrt{\epsilon} \cdot |S_2|}{c_1' \cdot m_{\sqrt{\epsilon},\delta}} \geq c_2'\rho_I(x) = \frac{c_2'\rho_2(x)}{2},$$

$$\rho(x) = \rho_I(x)\tfrac{1}{2}\sqrt{\epsilon}$$

Algorithm

Using the super-sampling guarantees of Lemma 2, with probability $1 - 2\delta/3$, $\mathrm{err}_{D_2}(h_2) \leq \sqrt{\epsilon}/2$. ■

Phase 3:

Let $\overline{S_3} = \textsc{Correct-Label}(S_3, \delta/6)$, for a sample set S_3 of size $2m_{\sqrt{\epsilon},\delta/6}$ drawn from $D_{|\mathcal{X}}$ conditioned on $h_1(x) \neq h_2(x)$.

Let $h_3 = \mathcal{O}_{\mathcal{F}}(\overline{S_3})$.

return $\mathrm{Maj}(h_1, h_2, h_3)$.

$$\text{with probability } 1-\delta, err_D(MAJ(h_1, h_2, h_3)) \leq \varepsilon$$

Algorithm

Lemma 9 *Let S be a sample set drawn from distribution D and let h be such that $\mathrm{err}_D(h) \leq \sqrt{\epsilon}$. With probability $1 - \exp(-\Omega(|S|\sqrt{\epsilon}))$, $\textsc{Filter}(S, h)$ makes $O(|S|)$ label queries.*

Proof of Theorem 3 We first discuss the number of label queries Algorithm 2 makes. The total number of labels queried by Phases 1 and 3 is attributed to the labels queried by $\textsc{Correct-Label}(S_1, \delta)$ and $\textsc{Correct-Label}(S_3, \delta/6)$, which is $O\left(m_{\sqrt{\epsilon},\delta} \log(m_{\sqrt{\epsilon},\delta}/\delta)\right)$. By Lemma 7, $|S_I \cup S_C| \leq O(m_{\sqrt{\epsilon},\delta})$ almost surely. So, $\textsc{Correct-Label}(S_I \cup S_C, \delta/6)$ contributes $O\left(m_{\sqrt{\epsilon},\delta} \log(m_{\sqrt{\epsilon},\delta}/\delta)\right)$ labels. Moreover, as we showed in Lemma 9, $\textsc{Filter}(S_2, h_1)$ queries $O(m_{\epsilon,\delta})$ labels, almost surely. So, the total number of labels queried by Algorithm 2 is at most $O\left(m_{\sqrt{\epsilon},\delta} \log\left(\frac{m_{\sqrt{\epsilon},\delta}}{\delta}\right) + m_{\epsilon,\delta}\right)$. This leads to $\Lambda = O\left(\sqrt{\epsilon} \log\left(\frac{m_{\sqrt{\epsilon},\delta}}{\delta}\right) + 1\right)$ cost per labeled example.

$$m_{c\varepsilon,\delta} = O(\frac{1}{c} m_{\varepsilon,\delta})$$

Algorithm

Theorem 3 ($\alpha = \frac{1}{2} + \Theta(1)$ **case**) *Algorithm 2 uses oracle* $\mathcal{O}_{\mathcal{F}}$, *runs in time* $\mathrm{poly}(d, \frac{1}{\epsilon}, \ln(\frac{1}{\delta}))$ *and with probability* $1 - \delta$ *returns* $f \in \mathcal{F}$ *with* $\mathrm{err}_D(f) \leq \epsilon$, *using* $\Lambda = O\left(\sqrt{\epsilon}\log\left(\frac{m_{\sqrt{\epsilon},\delta}}{\delta}\right) + 1\right)$ *cost per labeled example,* $\Gamma = 0$ *golden queries, and* $\lambda = 1$ *load. Note that when* $\frac{1}{\sqrt{\epsilon}} \geq \log\left(\frac{m_{\sqrt{\epsilon},\delta}}{\delta}\right)$, *the above cost per labeled sample is* $O(1)$.

Baseline algorithm takes $\log(\frac{m_{\varepsilon,\delta}}{\delta})$ more labels than it requires samples

Algorithm

When $\alpha < \frac{1}{2} + o(1), CORRECT - LABEL(S, \delta)$ and $FILTER(S,h)$ will make mistakes.

PRUNE-AND-LABEL(S, δ)**:**

---

for $x \in S$ **do**

Let $L \sim P^k$ for a set of $k = O(\frac{1}{\alpha^2} \log(\frac{|S|}{\delta}))$ labelers drawn from P.

if Maj-size$_L(x) \leq 1 - \frac{\alpha}{4}$ **then**

Get a golden query $y^* = f^*(x)$,

Restart Algorithm 3 with distribution $P \leftarrow P_{|\{(x,y^*)\}}$ and $\alpha \leftarrow \frac{\alpha}{1-\frac{\alpha}{8}}$.

else

$\overline{S} \leftarrow \overline{S} \cup \{(x, \text{Maj}_L(x))\}$.

end

end

return $\overline{S}$.

Algorithm

Algorithm 3 BOOSTING BY PROBABILISTIC FILTERING FOR ANY α

Input: Given a distribution $D_{|\mathcal{X}}$ and P, a class of hypothesis $\mathcal{F}$, parameters ϵ, δ, and α.

Phase 0:

If $\alpha > \frac{3}{4}$, run Algorithm 2 and quit.

Let $\delta' = c\alpha\delta$ for small enough $c > 0$ and draw S_0 of $O(\frac{1}{\epsilon}\log(\frac{1}{\delta'}))$ examples from the distribution D.

PRUNE-AND-LABEL(S_0, δ').

Phase 1:

Let $\overline{S_1}$ = PRUNE-AND-LABEL(S_1, δ'), for a set of sample S_1 of size $2m_{\sqrt{\epsilon},\delta'}$ from D.

Let $h_1 = \mathcal{O}_{\mathcal{F}}(\overline{S_1})$.

Algorithm

Phase 2:

Let $S_I = \text{FILTER}(S_2, h_1)$, for a set of samples S_2 of size $\Theta(m_{\epsilon,\delta'})$ drawn from D.

Let S_C be a sample set of size $\Theta(m_{\sqrt{\epsilon},\delta'})$ drawn from D.

Let $\overline{S_{All}} = \text{PRUNE-AND-LABEL}(S_I \cup S_C, \delta')$.

Let $\overline{W_I} = \{(x, y) \in \overline{S_{All}} \mid y \neq h_1(x)\}$ and Let $\overline{W_C} = \overline{S_{All}} \setminus \overline{W_I}$.

Draw a sample set $\overline{W}$ of size $\Theta(m_{\sqrt{\epsilon},\delta'})$ from a distribution that equally weights $\overline{W_I}$ and $\overline{W_C}$.

Let $h_2 = \mathcal{O}_{\mathcal{F}}(\overline{W})$.

Phase 3:

Let $\overline{S_3} = \text{PRUNE-AND-LABEL}(S_3, \delta')$, for a sample set S_3 of size $2m_{\sqrt{\epsilon},\delta'}$ drawn from D conditioned on $h_1(x) \neq h_2(x)$.

Let $h_3 = \mathcal{O}_{\mathcal{F}}(\overline{S_3})$.

return $\text{Maj}(h_1, h_2, h_3)$.

Algorithm

Lemma 11 *For any δ, with probability $1-\delta$, the total number of times that Algorithm 3 is restarted as a result of pruning is $O(\frac{1}{\alpha})$.*

Recall that $\delta' = c \cdot \alpha\delta$ for some small enough constant $c > 0$. Each time PRUNE-AND-LABEL(S, δ') is called, by Hoeffding bound, it is guaranteed that with probability $\geq 1 - \delta'$, for each $x \in S$,

$$|\text{Maj-size}_P(x) - \text{Maj-size}_L(x)| \leq \frac{\alpha}{8},$$

where L is the set of labelers PRUNE-AND-LABEL(S, δ') queries on x. Hence, when we issue a golden query for x such that $\text{Maj-size}_L(x) \leq 1 - \frac{\alpha}{4}$ and prune away bad labelers, we are guaranteed to remove at least an $\frac{\alpha}{8}$ fraction of the labelers. Furthermore, no good labeler is ever removed. Hence, the fraction of good labelers increases from α to $\alpha/(1 - \frac{\alpha}{8})$. So, in $O(\frac{1}{\alpha})$ calls, the fraction of the good labelers surpasses $\frac{3}{4}$ and we switch to using Algorithm 2. Therefore, with probability $1 - \delta$ overall, the total number of golden queries is $O(1/\alpha)$.

Algorithm

Lemma 2 *Given a hypothesis class $\mathcal{F}$ consider any two discrete distributions D and D' such that for all x, $\rho'(x) \geq c \cdot \rho(x)$ for an absolute constant $c > 0$, and both distributions are labeled according to $f^* \in \mathcal{F}$. There exists a constant $c' > 1$ such that for any ϵ and δ, with probability $1 - \delta$ over a labeled sample set S of size $c' m_{\epsilon,\delta}$ drawn from D', $\mathcal{O}_{\mathcal{F}}(S)$ has error of at most ϵ with respect to distribution D.*

Lemma 12 (Robust Super-Sampling Lemma) *Given a hypothesis class $\mathcal{F}$ consider any two discrete distributions D and D' such that except for an ϵ fraction of the mass under D, we have that for all x, $\rho'(x) \geq c \cdot \rho(x)$ for an absolute constant $c > 0$ and both distributions are labeled according to $f^* \in \mathcal{F}$. There exists a constant $c' > 1$ such that for any ϵ and δ, with probability $1 - \delta$ over a labeled sample set S of size $c' m_{\epsilon,\delta}$ drawn from D', $\mathcal{O}_{\mathcal{F}}(S)$ has error of at most 2ϵ with respect to D.*

Algorithm

Let B be the set of points that do not satisfy the condition that $\rho'(x) \geq c \cdot \rho(x)$. Notice that because D and D' are both labeled according to $f^* \in \mathcal{F}$, for any $f \in \mathcal{F}$ we have,

$$\mathrm{err}_{D'}(f) = \sum_{x \in B} \rho'(x) \mathbb{1}_{f(x) \neq f^*(x)} + \sum_{x \notin B} \rho'(x) \mathbb{1}_{f(x) \neq f^*(x)} \geq \sum_{x \notin B} c \cdot \rho(x) \mathbb{1}_{f(x) \neq f^*(x)} \geq c \cdot (\mathrm{err}_D(f) - \epsilon).$$

Therefore, if $\mathrm{err}_{D'}(f) \leq c\epsilon$, then $\mathrm{err}_D(f) \leq 2\epsilon$. Let $m' = m_{c\epsilon,\delta}$, we have

$$\begin{aligned} \delta &> \Pr_{S' \sim D'^{m'}} [\exists f \in \mathcal{F}, \text{s.t. } \mathrm{err}_{S'}(f) = 0 \wedge \mathrm{err}_{D'}(f) \geq c\epsilon] \\ &\geq \Pr_{S' \sim D'^{m'}} [\exists f \in \mathcal{F}, \text{s.t. } \mathrm{err}_{S'}(f) = 0 \wedge \mathrm{err}_{D}(f) \geq 2\epsilon]. \end{aligned}$$

The claim follows by the fact that $m_{c\epsilon,\delta} = O\left(\frac{1}{c} m_{\epsilon,\delta}\right)$.

Algorithm

Theorem 13 (Any α) *Suppose the fraction of the perfect labelers is α and let $\delta' = c\alpha\delta$ for small enough constant $c > 0$. Algorithm 3 uses oracle $\mathcal{O}_{\mathcal{F}}$, runs in time $\mathrm{poly}(d, \frac{1}{\alpha}, \frac{1}{\epsilon}, \ln(\frac{1}{\delta}))$, uses a training set of size $O(\frac{1}{\alpha} m_{\epsilon,\delta'})$ size and with probability $1-\delta$ returns $f \in \mathcal{F}$ with $\mathrm{err}_D(f) \leq \epsilon$ using $O(\frac{1}{\alpha})$ golden queries, load of $\frac{1}{\alpha}$ per labeler, and a total number of queries*

$$O\left(\frac{1}{\alpha} m_{\epsilon,\delta'} + \frac{1}{\alpha\epsilon}\log(\frac{1}{\delta'})\log(\frac{1}{\epsilon\delta'}) + \frac{1}{\alpha^3} m_{\sqrt{\epsilon},\delta'}\log(\frac{m_{\sqrt{\epsilon},\delta'}}{\delta'})\right).$$

Note that when $\frac{1}{\alpha^2\sqrt{\epsilon}} \geq \log\left(\frac{m_{\sqrt{\epsilon},\delta}}{\alpha\delta}\right)$ and $\log(\frac{1}{\alpha\delta}) < d$, the cost per labeled query is $O(\frac{1}{\alpha})$.

Algorithm

No perfect Labelers

Good labelers have error of $\leq \varepsilon$

Bad labelers have error of $\geq 4\varepsilon$

Assume that more than half of labelers are good

Two good labelers agree on at least $1-2\varepsilon$ fraction of the data

A bad and a good labeler agree on at most $1-3\varepsilon$ of the data

Algorithm

Algorithm 4 GOOD LABELER DETECTION

Input: Given n labelers, parameters ϵ and δ

Let $G = ([n], \emptyset)$ be a graph on n vertices with no edges.

Take set Q of $16\ln(2)n$ random pairs of nodes from G.

1 **for** $(i, j) \in Q$ **do**

  if DISAGREE$(i, j) < 2.5\epsilon$ **then** add edge (i, j) to G;

end

2 Let $\mathcal{C}$ be the set of connected components of G each with $\geq n/4$ nodes.

3 **for** $i \in [n] \setminus \left(\bigcup_{C \in \mathcal{C}} C\right)$ *and* $C \in \mathcal{C}$ **do**

  Take one node $j \in C$, if DISAGREE$(i, j) < 2.5\epsilon$ add edge (i, j) to G.

end

return *The largest connected component of* G

DISAGREE(i, j):

Take set S of $\Theta(\frac{1}{\epsilon}\ln(\frac{n}{\delta}))$ samples from D.

return $\frac{1}{|S|}\sum_{x \in S} \mathbb{1}_{(g_i(x) \neq g_j(x))}$.

Algorithm

Theorem 14 *Suppose that any* good *labeler* i *is such that* $\mathrm{err}_D(g_i) \leq \epsilon$. *Furthermore, assume that* $\mathrm{err}_D(g_j) \notin (\epsilon, 4\epsilon)$ *for any* $j \in [n]$. *And let the number of good labelers be at least* $\lfloor \frac{n}{2} \rfloor + 1$. *Then, Algorithm 4, returns the set of all good labeler with probability* $1 - \delta$, *using an expected load of* $\lambda = O\left(\frac{1}{\epsilon} \ln\left(\frac{n}{\delta}\right)\right)$ *per labeler.*