Partial Correction

Arushi Gupta ag3309

Columbia University

Nov 20

Outline

Recall active learning

Taxonomy

Threshold functions

Main algorithm

Stick with it

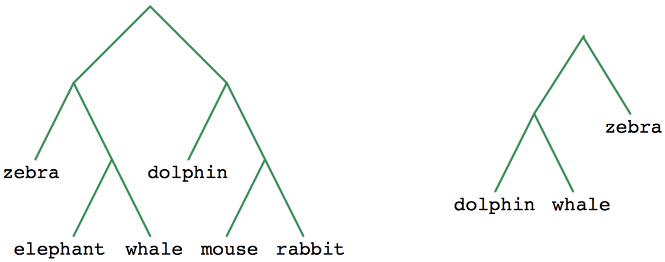
Introduction: recall active learning

- We have some distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$
- the set of hypothesis $\mathcal{H}$ maps $\mathcal{X}$ to $\mathcal{Y}$
- at time t we observe $x_t \in \mathcal{X}$ and decide where or not to query its label

Introduction

Taxonomy

In previous models of interactive learning (active learning) we asked a question and received an answer. But what if were trying to solve a more complex problem.



Introduction

- There exists a space of structures $\mathcal{H}$ (trees over species)
- some $q \in \mathcal{Q}$ is chosen at random
- the learner displays q and h(q) to some expert
- if h(q) is correct, the expert accepts it, otherwise the expert corrects some part of it

Examples

What do we mean by "part of it?" Assume q has c atomic components. We will discuss how the expert picks the component.

Introduction

- We will write $q \in_\mu \mathcal{Q}$ to indicate q was chosen according to probability distribution μ from $\mathcal{Q}$ and $[c] = \{1, 2, ..., c\}$
- How do we measure error?
  - by the full question q, i.e.

$$err(h) = P_{q \in_\mu \mathcal{Q}}[h(q) \neq h^*(q)] \tag{1}$$

  - in terms of components i.e.

$$err_c(h) = P_{q \in_\mu \mathcal{Q}, j \in_R [c]}[h(q, j) \neq h^*(q, j)] \tag{2}$$

Threshold functions

- let $\mathcal{X} = [0, 1]$
- let $\mathcal{H} = \{h_v : v \in [0, 1]\}$ and $h_v(x) = 1(x > v)$

Threshold functions

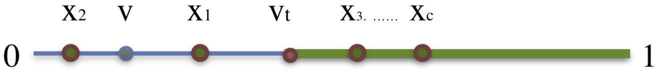
- Suppose we want to learn $h^* = h_0$
- our queries will consist of c numbers in $[0, 1]$ ($\mathcal{Q} = \mathcal{X}^c$)
- these numbers are our atomic components
- consider the uniform distribution μ on components.
- $err_c(h_v) = v \; err(h_v) = 1 - (1 - v)^c$

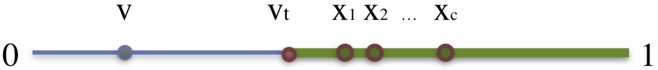
Threshold functions

- let v_t be the threshold learned so far by the algorithm
- labeling policy is "largest"
- labeling policy is "smallest"

Labeling policy is the largest

- let v_t be the threshold learned so far by the algorithm
- Let V_{t+1} be the random variable that is the threshold value the learner learns at step $t+1$
- pick a v in $[0, v_t)$. Then V_{t+1} can exceed v is if all pts are to the right of v_t. Or if there is a pt in (v, v_t)

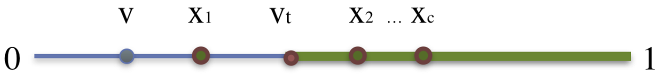




Labeling policy is the smallest

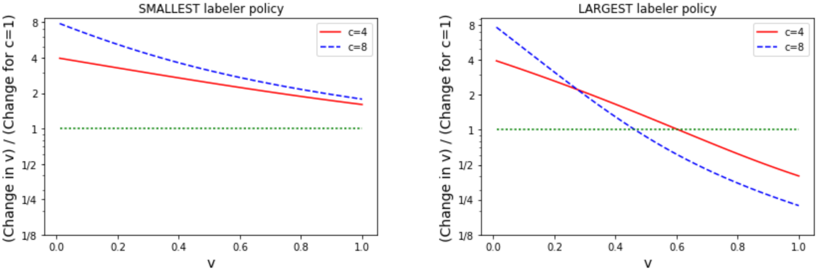
expectation

None of the x_i can lie in $[0, v]$



How does this compare to the largest labeling policy case? The improvement in the threshold is $\mathbb{E}[v_t - V_{t+1}]$

Labeling policy is the smallest



Threshold functions

- Suppose the support is on only $(1/c, 2/c, ..., c/c = 1)$, and suppose the expert corrects the most glaring error.
- it takes $c/2$ rounds to bring the error down to $1/2$

Different μ

- Suppose now that μ is supported on two points:

$$(\frac{1}{2c}, \frac{2}{2c}, ..., \frac{1}{2})$$

w.p. 2ϵ

$$(\frac{1}{2} + \frac{1}{2c}, \frac{1}{2} + \frac{2}{2c}, ..., 1)$$

w.p. $1 - 2\epsilon$

Say we want $err_c(h) \leq \epsilon$. We want

$$E_{q \in \mu \mathcal{Q}, j \in R[c]}[I_{h(q,j) \neq h_0(q,j)}] \leq \epsilon \tag{3}$$

. But h and h_0 will always agree on $[v, 1]$. So we want

$$P[\text{pick } x_i \in [0, v]] \leq \epsilon \Rightarrow v \leq 1/4 \tag{4}$$

So we must see the first pt at least $c/2$ times which requires $\Omega(c/\epsilon)$ examples.

Different μ

So we have shown

- Theorem 1. There is a concept class $\mathcal{H}$ of VC dimension 1 such that for any $\epsilon > 0$ it is necessary to have $O(c/\epsilon)$ rounds of feedback in order to be able to guarantee that with high prob all consistent hypotheses have error $\leq \epsilon$

Main result

- There exists a space of structures $\mathcal{H}$
- some $q \in_\mu \mathcal{Q}$ is chosen at random
- the learner displays q and h(q) to some expert
- if h(q) is correct, the expert accepts it, otherwise the expert corrects some part of it

main thm

Let $B(h) = \{q \in \mathcal{Q}$ s.t. h is incorrect on q $\}$ Let $G(h) = \{q \in \mathcal{Q}$ s.t. h is correct on q$\}$. The algorithm produce a hypothesis with error $\leq \epsilon$ w.p. at least $1 - \delta$ within $2N$ steps where $N = c \cdot (\frac{I}{\epsilon'} + 1)$. $I = \log(|\mathcal{H}|/\delta)$ and $\epsilon' = \epsilon/2$

Main result

- Let $\bar{\mathcal{Q}} = \mathcal{Q} \times [c]$
- $\bar{B}(h) = \{(q,j) \in \bar{\mathcal{Q}} : q \in B(h) \text{ and } h(q,j) \neq h^*(q,j)\}$
- $\bar{G}(h) = G(h) \times [c]$
- Let $\gamma(q,j)$ be the conditional probability that the expert provides feedback on j given that q is queried
- $w_t(q,j) = \mu(q) \cdot \gamma(q,j)$
- we are going calculate $w_t(q,1), ..., w_t(q,c)$ for $q \in G(h_t)$
- let $W_t(q,j) = w_1(q,j) + ... + w_t(q,j)$

How to pick the weights

Lemma 3

for all $q \in G(h_t)$ non negative values $w(q,1), ..., w(q,c)$ summing up to $\mu(q)$ can be calculated such that

$$W_t(q,j) = W_{t-1}(q,j) + w_t(q,j) \leq \frac{t \cdot \mu(q)}{c} \tag{5}$$

Proof

want to show

$$W_t(q,j) = W_{t-1}(q,j) + w_t(q,j) \leq \frac{t \cdot \mu(q)}{c} \tag{6}$$

Proof

$W_t(q,[c]) = t \cdot \mu(q)$. Pick $j_1, ..., j_c$ s.t.

$$W_{t-1}(q,j_1) \leq W_{t-1}(q,j_2) \leq ... \leq W_{t-1}(q,j_c) \tag{7}$$

Let $\Delta = \mu(q)$. initialize all the $w_t(q,j_i)$ to 0. repeat the following till $\Delta = 0$

$$w_t(q,j_i) = \min\{\frac{t \cdot \mu(q)}{c} - W_{t-1}(q,j_i), \Delta\} \tag{8}$$

and reset $\Delta = \Delta - w_t(q,j_i)$

Eliminating inconsistent hypotheses

main thm

With probability at least $1-\delta$, the following holds $\forall h \in \mathcal{H}$: If there is a step t for which $W_t(\bar{B}(h)) \geq l$, then h is not consistent with the feedback received up to that step

- any $h \in \mathcal{H}$ is eliminated w.p. at least $w_t(\bar{B}(h))$
- let t be the first step for which $W_t(\bar{B}(h)) \geq l$. Then the probability that h is not eliminated by the end of step t is

$$\begin{aligned}\left(1-w_1(\bar{B}(h))\right)\cdot\left(1-w_2(\bar{B}(h))\right)\cdots\left(1-w_t(\bar{B}(h))\right) \\ \leq \exp(-W_t(\bar{B}(h))) \\ \leq \frac{\delta}{|\mathcal{H}|}\end{aligned} \tag{9}$$

- now take the union bound over $\mathcal{H}$

Analyzing the first N steps

analysis

Let $\tau = \frac{N}{c} = \frac{l}{\epsilon'} + 1$ be a threshold value. We will think of an atomic component as having been adequately sampled when W_t reaches $\tau \cdot \mu(q)$. At the beginning of step t let
$\bar{L}_{t-1} = \{(q,j) \in \bar{\mathcal{Q}} : W_{t-1}(q,j) \leq \tau \cdot \mu(q)\}$ and let
$W_{t-1}(\bar{L}_{t-1}) = \sum_{(q,j)\in \bar{L}_{t-1}} W_{t-1}(q,j) \leq c \cdot \tau = N$ finally let
$\bar{L}'_{t-1} = \{(q,j) \in \bar{\mathcal{Q}} : W_{t-1}(q,j) \leq (\tau - 1) \cdot \mu(q) = \frac{l}{\epsilon'} \cdot \mu(q)\}$

lemma 5

previous definitions

$\bar{L}'_{t-1} = \{(q,j) \in \bar{\mathcal{Q}} : W_{t-1}(q,j) \leq (\tau - 1) \cdot \mu(q) = \frac{l}{\epsilon'} \cdot \mu(q)\}$

Statement

at any step t if $W_{t-1}(\bar{B}(h_t)) < l$ then

$$w_t(\bar{B}(h_t) \cap \bar{L}'_{t-1}) \geq \mu(B(h_t)) - \epsilon' \tag{10}$$

proof

Note that
$\mu(B(h_t)) = w_t(\bar{B}(h_t)) = w_t(\bar{B}(h_t) \cap L'_{t-1}) + w_t(\bar{B}(h_t) \setminus L'_{t-1})$.
Then we can see that

$$l > W_{t-1}(\bar{B}(h_t)) \geq W_{t-1}(\bar{B}(h_t) \setminus \bar{L}'_{t-1}) \geq \frac{l}{\epsilon'} \cdot w_t(\bar{B}(h_t) \setminus \bar{L}'_{t-1}) \tag{11}$$

. It follows that $w_t(\bar{B}(h_t) \setminus \bar{L}'_{t-1}) \leq \epsilon'$

Lemma 6

previous definitions

$\tau = \frac{N}{c} = \frac{l}{\epsilon'} + 1$
$\bar{L}_{t-1} = \{(q,j) \in \bar{\mathcal{Q}} : W_{t-1}(q,j) \leq \tau \cdot \mu(q)\}$
$\bar{L}'_{t-1} = \{(q,j) \in \bar{\mathcal{Q}} : W_{t-1}(q,j) \leq (\tau - 1) \cdot \mu(q) = \frac{l}{\epsilon'} \cdot \mu(q)\}$

Statement

at any step $t \leq N$, $w_t(\bar{L}_t) \geq 1 - \epsilon'$

proof

note that $w_t(\bar{L}_t) = w_t(\bar{B}(h_t) \cap \bar{L}_t) + w_t(\bar{G}(h_t) \cap \bar{L}_t)$. Since any $(q,j) \in \bar{B}(h_t) \cap \bar{L}'_{t-1}$ satisfies $(q,j) \in \bar{B}(h_t) \cap \bar{L}_t$ the previous lemma 5 implies $w_t(\bar{B}(h_t) \cap \bar{L}_t) \geq \mu(B(h_t)) - \epsilon'$. For $q \in G(h_t)$ any (q,j) with $w_t(q,j) > 0$ satisfies

$$W_t(q,j) \leq \frac{t \cdot \mu(q)}{c} \leq \tau \cdot \mu(q). \tag{12}$$

Lemma 6 continued

Statement

at any step $t \leq N$, $w_t(\bar{L}_t) \geq 1 - \epsilon'$

proof

Thus $(q, j) \in \bar{L}_t$ and it follows that

$$w_t(\bar{G}(h_t) \cap \bar{L}_t) = \mu(G(h_t)) \tag{13}$$

. Overall,

$$w_t(\bar{L}_t) \geq \mu(B(h_t)) - \epsilon' + \mu(G(h_t)) = 1 - \epsilon' \tag{14}$$

Corollary

Definitions

$\bar{L}_{t-1} = \{(q,j) \in \bar{\mathcal{Q}} : W_{t-1}(q,j) \leq \tau \cdot \mu(q)\}$ and let $W_{t-1}(\bar{L}_{t-1}) = \sum_{(q,j)\in \bar{L}_{t-1}} W_{t-1}(q,j) \leq c \cdot \tau = N$

Previous fact

$\forall t \leq N \; w_t(\bar{L}_t) \geq 1 - \epsilon'$

analysis

Let $\hat{W}_t(q,j) = \min\{W_t(q,j), \tau \cdot \mu(q)\}$. As we have seen, $\hat{W}_t(\bar{Q}) \leq N$. We can see as a corollary to before that $\hat{W}_N(\bar{Q}) \leq (1-\epsilon')N$.

Next N steps

analysis

Say $\mu(B(h_t)) \geq 2\epsilon'$. Then $\mu(B(h_t)) - \epsilon' \geq \epsilon'$ During one of the steps in the second phase $\mu(B(h_t)) < 2 \cdot \epsilon' = \epsilon$ at which point the algorithm can return h_t

Stick with it algorithm

analysis

There are some problems with the algorithm we described.

- you need to select a hypothesis that is consistent with feedback so far
- if you want an algorithm that is verified to have error less than ϵ you would need to run a separate procedure
- What if $|\mathcal{H}|$ is unbounded but the VC dimension is bounded?

Stick with it algorithm

- when you pick a hypothesis, stick with it for k steps.
- Redefine $N = c \cdot (\frac{l}{\epsilon'} + k)$. All parameters defined in terms of n are similarly defined.
- redefine $\bar{L}'_t = \{(q, j) \in \bar{Q} : W_t(q, j) \leq (\tau - k)\mu(q) = \frac{l}{\epsilon'} \cdot \mu(q)\}$

Then we have that

Stick with it algorithm

- The algorithm terminates in $2 \cdot N$ steps as before
- we can now use the k steps to verify the hypothesis
- we can define $l = d + \log(1/\delta)$ where d is the VC dimension of $\mathcal{H}$...where did we use this again?

main thm

With probability at least $1-\delta$, the following holds $\forall h \in \mathcal{H}$: If there is a step t for which $W_t(\bar{B}(h)) \geq I$, then h is not consistent with the feedback received up to that step

- any $h \in \mathcal{H}$ is eliminated w.p. at least $w_t(\bar{B}(h))$
- let t be the first step for which $W_t(\bar{B}(h)) \geq I$. Then the probability that h is not eliminated by the end of step t is

$$\begin{aligned}\left(1-w_1(\bar{B}(h))\right)\cdot\left(1-w_2(\bar{B}(h))\right)\cdots\left(1-w_t(\bar{B}(h))\right) \\ \leq \exp(-W_t(\bar{B}(h))) \\ \leq \frac{\delta}{|\mathcal{H}|}\end{aligned} \tag{15}$$

- now take the union bound over $\mathcal{H}$