Some Results from “Learning in the Presence of Malicious Errors”, 88’ STOC, Michael Kearns and Ming Li

Mengfei Cao

Tufts University

mcao01@cs.tufts.edu

March 5, 2013

Good Times

The original PAC learning model

X: instance space
C: concept class
H: hypothesis class
POS: an oracle that gives a positive example in unit time
NEG: an oracle that gives a negative example in unit time
D^+: distribution over the positive subset of X given a concept c
D^-: distribution over the negative subset of X given a concept c

Definition

We say C is PAC-learnable by H over X if:
$\exists$ algorithm A, s.t.
$\forall \epsilon, \delta$: the input from $(0, 1)$
$\forall c \in C$: the target concept,
$\forall D^+, D^-$: instance distribution w.r.t. c,
$h = A(\epsilon, \delta)$: $err^+(h) < \epsilon$ and $err^-(h) < \epsilon$ with prob. at least $1 - \delta$
by accessing POS and NEG and running in finite amount of steps
where $err^+ = D_c^+(neg(h))$ and $err^- = D_c^-(pos(h))$

One More Thing

We have proved that 2-oracle model is equivalent to 1-oracle model

Not-So-Good Times

Learning with Malicous Errors

X: instance space

C: concept class

H: hypothesis class

$$POS_{MAL}^{\beta} : \begin{cases} POS_{old} & \text{w.p. } 1-\beta \\ \textit{some adversary} & \text{w.p. } \beta \end{cases}$$

$$NEG_{MAL}^{\beta} : \begin{cases} NEG_{old} & \text{w.p. } 1-\beta \\ \textit{some adversary} & \text{w.p. } \beta \end{cases}$$

D^+: distribution over the positive subset of X given a concept c

D^-: distribution over the negative subset of X given a concept c

Not-So-Good Times

β-tolerant PAC-Learning ($0 \leq \beta < 1/2$)

We say C is β-tolerant PAC-learnable by H over X if:
$\exists$ algorithm A, s.t.
$\forall \epsilon, \delta$: the input from $(0, 1)$
$\forall c \in C$: the target concept,
$\forall D^+, D^-$: instance distribution w.r.t. c,
$h = A(\epsilon, \delta)$: $err^+(h) < \epsilon$ and $err^-(h) < \epsilon$ with prob. at least $1 - \delta$
by accessing POS^{β}_{MAL} and NEG^{β}_{MAL} and running in finite amount of steps
where $err^+ = D^+_c(neg(h))$ and $err^- = D^-_c(pos(h))$

How “Not-So-Good” Can “Oracles” Be?

If POS^{β}_{MAL} and NEG^{β}_{MAL} always behave strangely

What is the largest possible β so that we can still learn concepts?

Why do we have $\beta < 1/2$?

How "Not-So-Good" Can "Oracles" Be?

If POS^{β}_{MAL} and NEG^{β}_{MAL} always behave strangely

for example, when $\beta = 1$ and the adversary makes all "concepts" look like the same by manipulating examples. We can't learn correct concepts, not even close.

What is the largest possible β so that we can still learn concepts?

Why do we have $\beta < 1/2$?

How “Not-So-Good” Can “Oracles” Be?

If POS^{β}_{MAL} and NEG^{β}_{MAL} always behave strangely

for example, when $\beta = 1$ and the adversary makes all “concepts” look like the same by manipulating examples. We can't learn correct concepts, not even close.

What is the largest possible β so that we can still learn concepts?

A: PAC learning algorithm for C
$E_{MAL}(C, A)$: defined to be the largest β such that A is a β-tolerant learning algorithm for C ($\sim(\epsilon, \delta, \beta)$)
$E_{MAL}(C)$: the supremum of $E_{MAL}(C, A)$ over all possible A

Why do we have $\beta < 1/2$?

How “Not-So-Good” Can “Oracles” Be?

If POS^{β}_{MAL} and NEG^{β}_{MAL} always behave strangely

for example, when $\beta = 1$ and the adversary makes all “concepts” look like the same by manipulating examples. We can't learn correct concepts, not even close.

What is the largest possible β so that we can still learn concepts?

A: PAC learning algorithm for C
$E_{MAL}(C, A)$: defined to be the largest β such that A is a β-tolerant learning algorithm for C ($\sim(\epsilon, \delta, \beta)$)
$E_{MAL}(C)$: the supremum of $E_{MAL}(C, A)$ over all possible A

Why do we have $\beta < 1/2$?

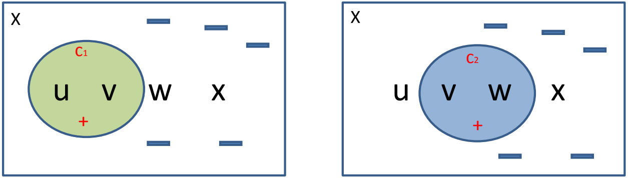
We will prove it for concept classes that are distinct

An Upper Bound for $E_{MAL}(C)$

Theorem

Definition

A concept class C is distinct iff $\exists c_1, c_2 \in C, u, v, w, x \in X$ s.t.



An Upper Bound for $E_{MAL}(C)$

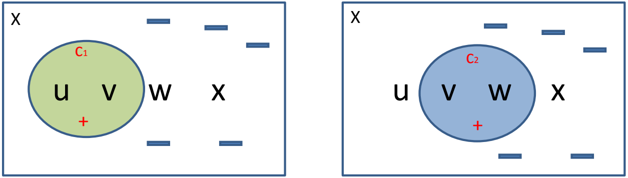
Theorem

Let C be a distinct representation class. Then

$$E_{MAL}(C) < \frac{\epsilon}{1+\epsilon}$$

Definition

A concept class C is distinct iff $\exists c_1, c_2 \in C, u, v, w, x \in X$ s.t.



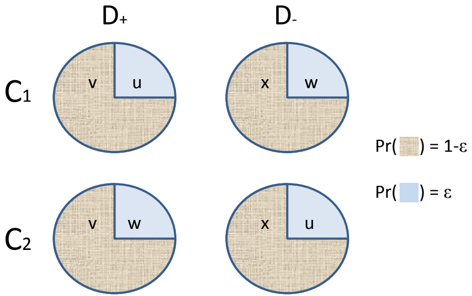
Proof for Theorem 1

Ideas:

Because C is distinct, we can make use of c_1, c_2, u, v, w, x and construct $D_1^+, D_1^-, D_2^+, D_2^-$ such that when β is at least $\frac{\epsilon}{1+\epsilon}$, c_1 and c_2 can't be learned for such distributions.

Proof:

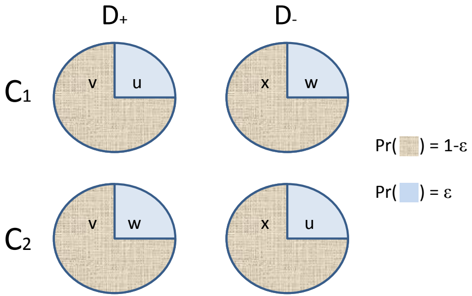
Construct D^+, D^- as follows:



Proof for Theorem 1, Cont'd

Worst-case Oracle:

Construct the adversary for c_1 as follows:
Whenever an error occurs (with prob. β), POS^{β}_{MAL} returns w and NEG^{β}_{MAL} returns u;
Construct the adversary for c_2 as follows:
Whenever an error occurs (with prob. β), POS^{β}_{MAL} returns u and NEG^{β}_{MAL} returns w;



Proof for Theorem 1, Cont'd

Induced Distribution:

When the target concept is c_1, if we access POS_{MAL}^{β}:

$Pr_{c_1}^{+}(u) = (1-\beta)\epsilon$, $Pr_{c_1}^{+}(v) = (1-\beta)(1-\epsilon)$, $Pr_{c_1}^{+}(w) = \beta$

if we access NEG_{MAL}^{β}:

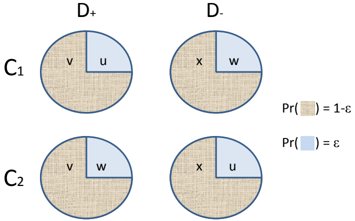
$Pr_{c_1}^{-}(w) = (1-\beta)\epsilon$, $Pr_{c_1}^{-}(x) = (1-\beta)(1-\epsilon)$, $Pr_{c_1}^{-}(u) = \beta$

When the target concept is c_2, if we access POS_{MAL}^{β}:

$Pr_{c_2}^{+}(u) = \beta$, $Pr_{c_2}^{+}(v) = (1-\beta)(1-\epsilon)$, $Pr_{c_2}^{+}(w) = (1-\beta)\epsilon$

if we access NEG_{MAL}^{β}:

$Pr_{c_2}^{-}(w) = \beta$, $Pr_{c_2}^{-}(x) = (1-\beta)(1-\epsilon)$, $Pr_{c_2}^{-}(u) = (1-\beta)\epsilon$



Proof for Theorem 1, Cont'd

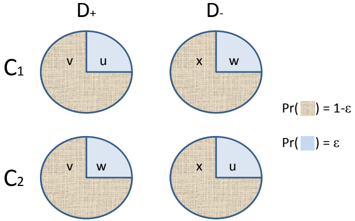
Induced Distribution:

When the target concept is c_1, if we access POS^{β}_{MAL}:

$Pr^{+}_{c_1}(u) = (1-\beta)\epsilon$, $Pr^{+}_{c_1}(v) = (1-\beta)(1-\epsilon)$, $Pr^{+}_{c_1}(w) = \beta$

When the target concept is c_2, if we access POS^{β}_{MAL}:

$Pr^{+}_{c_2}(u) = \beta$, $Pr^{+}_{c_2}(v) = (1-\beta)(1-\epsilon)$, $Pr^{+}_{c_2}(w) = (1-\beta)\epsilon$



Proof for Theorem 1, Cont'd

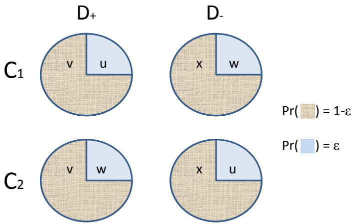
Induced Distribution:

When the target concept is c_1,

if we access NEG^{β}_{MAL}:
$Pr^{-}_{c_1}(w) = (1-\beta)\epsilon$, $Pr^{-}_{c_1}(x) = (1-\beta)(1-\epsilon)$, $Pr^{-}_{c_1}(u) = \beta$

When the target concept is c_2,

if we access NEG^{β}_{MAL}:
$Pr^{-}_{c_2}(w) = \beta$, $Pr^{-}_{c_2}(x) = (1-\beta)(1-\epsilon)$, $Pr^{-}_{c_2}(u) = (1-\beta)\epsilon$



Proof for Theorem 1, Cont'd

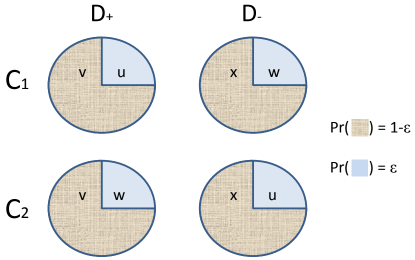
If $\beta = \epsilon/(1+\epsilon)$, then both the two pairs of distributions $(Pr^+_{c_1}(), Pr^+_{c_2}())$, $(Pr^-_{c_1}(), Pr^-_{c_2}())$ are identical respectively.

In other words, when we try to learn c_1 and c_2, oracles will give us examples from exactly the same distributions.

Even when $\beta > \epsilon/(1+\epsilon)$, the adversary can simulate D^+ and D^- appropriately so as to reduce the outcome error probability to $\epsilon/(1+\epsilon)$.

Proof for Theorem 1, Cont'd

Further



If h is ϵ-good hypothesis learnt for c_2, then
$err^+_{c_2}(h) = D^+_{c_2}(neg(h)) < \epsilon$,
$err^-_{c_2}(h) = D^-_{c_2}(pos(h)) < \epsilon$,
so $w \in pos(h)$ and $u \in neg(h)$. Yet:
$err^+_{c_1}(h) = D^+_{c_1}(neg(h)) \geq D^+_{c_1}(\{u\}) = \epsilon$
$err^-_{c_1}(h) = D^-_{c_1}(pos(h)) \geq D^-_{c_1}(\{w\}) = \epsilon$
Thus any ϵ-good hypothesis learnt for c_2 is ϵ-bad for c_1, vice versa.

Proof for Theorem 1, Cont'd

Therefore,

Concepts c_1 and c_2 can't both be learnt by any algorithms. Thus, C is not learnable if $\beta \geq \epsilon/(1+\epsilon)$

In all,

$$E_{MAL}(C) < \frac{\epsilon}{1+\epsilon}$$

, where C is a distinct concept class.

A Lower Bound for $E_{MAL}(C)$ & Sample Complexity Bound

Theorem

Definition

If an algorithm A accesses POS^{β}_{MAL} and NEG^{β}_{MAL} and takes inputs $0 < \epsilon, \delta < 1$; suppose that for target representation $c \in C$ and $0 \leq \beta < \epsilon/4$, A makes m calls to POS^{β}_{MAL} and recieves points $u_1, ..., u_m \in X$, and m calls to NEG^{β}_{MAL} and receives points $v_1, ..., v_m \in X$, and outputs $h_A \in H$ satisfying with probability at least $1 - \delta$, h_A is almost-consistent with positive sample and almost-consistent with negative sample, where "almost-consistent":
$|\{u_i : u_i \in neg(h_A)\}| \leq \frac{\epsilon}{2}m$ (for positive sample),
$|\{v_i : v_i \in pos(h_A)\}| \leq \frac{\epsilon}{2}m$ (for negative sample).
Such an algorithm A is a β-tolerant Occam algorithm for C by H

A Lower Bound for $E_{MAL}(C)$ & Sample Complexity Bound

Theorem

Let $\beta < \epsilon/4$, and A be a β-tolerant Occam algorithm for C by H. Then A is a β-tolerant learning algorithm for C by H; the sample size required is $m = O(1/\epsilon \ln 1/\delta + 1/\epsilon \ln |H|)$.

Definition

If an algorithm A accesses POS^{β}_{MAL} and NEG^{β}_{MAL} and takes inputs $0 < \epsilon, \delta < 1$; suppose that for target representation $c \in C$ and $0 \leq \beta < \epsilon/4$, A makes m calls to POS^{β}_{MAL} and recieves points $u_1, ..., u_m \in X$, and m calls to NEG^{β}_{MAL} and receives points $v_1, ..., v_m \in X$, and outputs $h_A \in H$ satisfying with probability at least $1 - \delta$, h_A is almost-consistent with positive sample and almost-consistent with negative sample, where "almost-consistent":
$|\{u_i : u_i \in neg(h_A)\}| \leq \frac{\epsilon}{2} m$ (for positive sample),
$|\{v_i : v_i \in pos(h_A)\}| \leq \frac{\epsilon}{2} m$ (for negative sample).
Such an algorithm A is a β-tolerant Occam algorithm for C by H

Proof for the Second Theorem

For simplicity, we prove for positive examples and the case for negative examples is similar.

Define bad hypothesis:

Fix a bad hypothesis h:

The prob. that h_{bad} is almost-consistent with positive sample:

Among $|H|$ hypothesis, the prob. that one such h_{bad} exists:

Proof for the Second Theorem

For simplicity, we prove for positive examples and the case for negative examples is similar.

Define bad hypothesis:

Let $h \in H$ be such that $e^{+}(h) \geq \epsilon$.

Fix a bad hypothesis h:

The prob. that h_{bad} is almost-consistent with positive sample:

Among $|H|$ hypothesis, the prob. that one such h_{bad} exists:

Proof for the Second Theorem

For simplicity, we prove for positive examples and the case for negative examples is similar.

Define bad hypothesis:

Let $h \in H$ be such that $e^+(h) \geq \epsilon$.

Fix a bad hypothesis h:

The probability that h agrees with a point received from POS_{MAL}^{β} :
$Pr(agree/no\ error) \cdot (1-\beta) + Pr(agree/error) \cdot \beta$
$\leq (1-\epsilon) \cdot (1-\beta) + \beta = 1 - \epsilon + \epsilon \cdot \beta \leq 1 - \epsilon + \epsilon/4 = 1 - \frac{3\epsilon}{4}$

The prob. that h_{bad} is almost-consistent with positive sample:

Among $|H|$ hypothesis, the prob. that one such h_{bad} exists:

Proof for the Second Theorem

For simplicity, we prove for positive examples and the case for negative examples is similar.

Define bad hypothesis:

Let $h \in H$ be such that $e^{+}(h) \geq \epsilon$.

Fix a bad hypothesis h:

The probability that h agrees with a point received from POS_{MAL}^{β} :
$Pr(agree/no\ error) \cdot (1-\beta) + Pr(agree/error) \cdot \beta$
$\leq (1-\epsilon) \cdot (1-\beta) + \beta = 1 - \epsilon + \epsilon \cdot \beta \leq 1 - \epsilon + \epsilon/4 = 1 - \frac{3\epsilon}{4}$

The prob. that h_{bad} is almost-consistent with positive sample:

Among m events of which each succeeds with prob. at least $\frac{3\epsilon}{4}$, at most $\epsilon/2$ happens. By Chernoff bounds, we have $\leq e^{-m\epsilon/24}$.

Among $|H|$ hypothesis, the prob. that one such h_{bad} exists:

Proof for the Second Theorem

For simplicity, we prove for positive examples and the case for negative examples is similar.

Define bad hypothesis:

Let $h \in H$ be such that $e^{+}(h) \geq \epsilon$.

Fix a bad hypothesis h:

The probability that h agrees with a point received from POS_{MAL}^{β} :
$Pr(agree/no\ error) \cdot (1 - \beta) + Pr(agree/error) \cdot \beta$
$\leq (1 - \epsilon) \cdot (1 - \beta) + \beta = 1 - \epsilon + \epsilon \cdot \beta \leq 1 - \epsilon + \epsilon/4 = 1 - \frac{3\epsilon}{4}$

The prob. that h_{bad} is almost-consistent with positive sample:

Among m events of which each succeeds with prob. at least $\frac{3\epsilon}{4}$, at most $\epsilon/2$ happens. By Chernoff bounds, we have $\leq e^{-m\epsilon/24}$.

Among $|H|$ hypothesis, the prob. that one such h_{bad} exists:

By union bound, the probability that one such hypothesis exists is at most $|H|e^{-m\epsilon/24}$. Solve $|H|e^{-m\epsilon/24} \leq \delta/2$ and we get $m \geq 24/\epsilon(\ln|H| + \ln 2/\delta)$.

Proof for the Second Theorem

The same argument also holds for NEG^{β}_{MAL}.

Thus,

if the output h is almost-consistent with both positive sample and negative sample, then with probability at least $1 - \delta$, the error probability is at most ϵ on both D^{+} and D^{-}, as long as $m \geq 24/\epsilon(\ln|H| + \ln 2/\delta)$.

Discussion

Efficiency

The second theorem gives a polynomial upper bound on the sample complexity for finite representation class ($|H|$ is finite), as well as an exhaustic search algorithm that is β-tolerant learning algorithm. However, the time complexity of such algorithm can be super-polynomial.

A tight bound on $E_{MAL}(C)$

Theorem 1 tells us that $E_{MAL}(C) < \frac{\epsilon}{1+\epsilon} = O(\epsilon)$ for distinct concept class. The second theorem tells us for finite representation class, that $\forall \beta < \epsilon/4$, C is efficiently learnable; in other words, $E_{MAL}(C) \geq \epsilon/4 = \Omega(\epsilon)$.
In conclusion, these give us the tight bound $\Theta(\epsilon)$ on $E_{MAL}(C)$ for distinct and finite representation class.

Endding

Practically,

No matter how "Not-So-Good" the oracles are, we can learn probably approximately correct concept given any accuracy parameter ϵ as long as the error probability β is stringently bounded by ϵ.

Thank you!

The End