BOOSTING

RECALL PAC LEARNING

$\underline{A}$

- $\underline{\delta}$ ← PROBABILITY OF FAILURE

- $\underline{\epsilon}$ ← ACCURACY PARAMETER

REQUIREMENT:

FOR ANY CHOICE OF ϵ, δ A SHOULD OUTPUT WITH PROBABILITY $\geq 1-\delta$, AN ϵ-ACCURATE CLASSIFIER.

A IS ALLOWED TO RUN IN TIME $\text{poly}\left(\frac{1}{\epsilon}, \frac{1}{\delta}\right)$

TAKE # OF SAMPLES $\text{poly}\left(\frac{1}{\epsilon}, \frac{1}{\delta}\right)$.

QUESTION: WHAT IF WE HAVE AN ALGORITHM A THAT WITH PROBABILITY 5% OUTPUTS AN $\underline{\epsilon}$-ACCURATE CLASSIFIER. HOW CAN WE USE A TO OBTAIN A STANDARD PAC LEARNER?

WE WANT TO INCREASE THAT 5% PROB OF SUCCESS TO $1-\delta$.

THE SOLUTION IS TO RUN A A LARGE # OF TIMES SAY t.

$$\Pr_{\text{run } A\ t \text{ times}}[A \text{ fails to output an } \epsilon\text{-ACCURATE CLASSIFIER}] \le (0.95)^t$$

WE CAN MAKE $(0.95)^t$ VERY SMALL BY CHOOSING t to be $\approx O(\log \frac{1}{\delta})$ THEN WE CAN "TEST" EACH CLASSIFIER GENERATED DURING THESE t trials to see if any OF THEM ARE GOOD CLASSIFIERS.

BOTTOM LINE: AMPLIFYING THE PROBABILITY OF SUCCESS IS NOT TOO DIFFICULT. $5\% \rightarrow 1-\delta$

TRICKIER QUESTION: WHAT IF ϵ IS FIXED TO SAY .49.

IMAGINE A WITH PROBABILITY $\geq 1-\delta$ OUTPUTS A CLASSIFIER WITH $\epsilon = .49$.

NATURAL QUESTION: HOW DO WE AMPLIFY/IMPROVE THE ACCURACY PARAMETER?

$\boxed{A} \to h \qquad err(h) = .49$

<u>WANT</u>

$$\begin{tikzcd} \boxed{A'} \arrow[r] \arrow[d] & h' \\ \boxed{A} & \end{tikzcd} \qquad err(h') \le \epsilon.$$

SOLUTION ATTEMPT: TRY RUNNING A MANY TIMES $h_1, \dots, h_t$

t

TAKE MAJ$(h_1, \dots, h_t)$

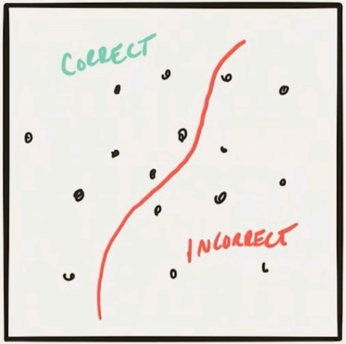
QUESTION WAS POSED BY VALIANT ON PAC LEARNING

SOLVED BY R. SCHAPIRE.

SOLUTION DUE TO FREUND & SCHAPIRE "BOOSTING ALGORITHMS"

ALGORITHM I'LL PRESENT: ADABOOST

HIGH-LEVEL IDEA



TRAINING SET

A outputs

h $\epsilon = .49$

FRACTION OF CORRECT POINTS IS .51

RUN A ON UNIFORM DIST ON POINTS IN TRAINING SET.

QUESTION WAS POSED BY VALIANT ON PAC LEARNING

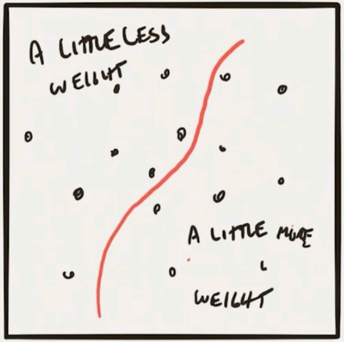
SOLVED BY R. SCHAPIRE.

"BOOSTING ALGORITHMS"

SOLUTION DUE TO FREUND & SCHAPIRE

ALGORITHM I'LL PRESENT: ADABOOST

HIGH-LEVEL IDEA



TRAINING SET

A outputs

h $\epsilon = .49$

FRACTION OF CORRECT POINTS IS .51

RUN A ON UNIFORM DIST ON POINTS IN TRAINING SET.

CORE IDEA:

- RE-WEIGHT POINTS WE GET WRONG TO HAVE MORE WEIGHT
- POINTS WE GET RIGHT: HAVE LESS WEIGHT
- RUN A AGAIN TO OBTAIN CLASSIFIER W.R.T. NEW WEIGHTING

MAJ(CLASSIFIERS GENERATED DURING THIS PROCESS)

ADABOOST (SIMPLIFIED VERSION)

- TRAINING SET OF SIZE m

INITIALLY D_0 = UNIFORM DIST. CORRESPONDS TO $w_i = 1 \ \forall i$

DIST IS OBTAINED BY DIVIDING BY W, SUM OF WEIGHTS.

E = error rate $\quad A$ = ACCURACY $= 1-E \quad \beta = \frac{E}{A}$

CONCRETELY $E = \frac{1}{2} - \gamma \qquad \beta = \dfrac{\frac{1}{2}-\gamma}{\frac{1}{2}+\gamma}$

HOW TO UPDATE WEIGHTS: AT ITERATION t, RUN A to OBTAIN $\underline{h_t}$

FOR EACH x_i s.t. $h_t(x_i)$ IS CORRECT $\quad w_i^{NEW} = \beta \cdot w_i^{OLD}$

" IS INCORRECT $\quad w_i^{NEW} = w_i^{OLD}$

REPEAT FOR T STEPS OUTPUT $MAJ(h_1, \ldots, h_T)$.

CLAIM: AFTER T ITERATIONS error $h_{FINAL} = MAJ(h_1, \ldots, h_T)$

$\leq e^{-2T\gamma^2}$ $\Rightarrow$ CHOOSE $T \approx \frac{1}{\gamma^2} \lg\left(\frac{1}{\epsilon}\right)$

then error of $h_{FINAL} \leq \epsilon$.

TOTAL
WEIGHT AFTER AN ITERATION

W IS WT OF ALL POINTS BEFORE ITERATION t

WT OF CORRECT POINTS AFTER ITERATION $t = \left(\frac{1}{2}+\gamma\right)\cdot\beta\cdot W$

" INCORRECT POINTS AFTER ITERATION $t = \left(\frac{1}{2}-\gamma\right)\cdot W$

RECALL $\left(\beta = \dfrac{\frac{1}{2}-\gamma}{\frac{1}{2}+\gamma}\right)$

NEW SUM OF ALL WEIGHTS? $W\left(\frac{1}{2}\beta + \gamma\beta + \frac{1}{2} - \gamma\right)$

$$W\left(\left(\frac{1}{2}+\gamma\right)\cdot\beta + \frac{1}{2} - \gamma\right)$$

$$W(1-2\gamma) = W\cdot\left(2\cdot\left(\frac{1}{2}-\gamma\right)\right).$$

AFTER i ITERATIONS SUM OF WEIGHTS $= W\cdot\left(2\left(\frac{1}{2}-\gamma\right)\right)^i$

$\Rightarrow$ AFTER T ITERATIONS SUM OF ALL WEIGHTS $\leq \underline{\left(\beta \cdot \left(\frac{1}{2} - \gamma\right)\right)^{T} \cdot W_0}$

CONSIDER A POINT X_i THAT h_{FINAL} GETS WRONG

$$wt(X_i) \geq \beta^{T/2}$$

$\Rightarrow$ IF h_{FINAL} HAS ERROR ϵ, THEN WT OF POINTS h_{FINAL} MISCLASSIFIES $\geq \underline{\epsilon \cdot m \cdot \beta^{T/2}}$

$$\underline{\epsilon \cdot \cancel{m} \cdot \beta^{\frac{T}{2}}} \le \underline{\left(2 \cdot \underbrace{\left(\tfrac{1}{2} - \gamma\right)}_{E}\right)^{T} \cdot \cancel{m}}$$

$$\epsilon \le \left(\frac{4 \cdot E^2}{\beta}\right)^{\frac{T}{2}}$$

$$\epsilon \le \left(1 - 4\gamma^2\right)^{\frac{T}{2}} \qquad \left(1 + x \approx e^{x}\right)$$

$$\le e^{-2\gamma^2 \cdot T}$$

IN ADABOOST

$h_t \in \{-1, +1\}$

$$\beta_t = \frac{E_t}{A_t}$$

OUTPUT: $\text{SIGN}\left(\sum_t \alpha_t h_t - \frac{1}{2}\right)$ $\quad \alpha_t = \dfrac{\log\left(\frac{1}{\beta_t}\right)}{\sum_t \log\left(\frac{1}{\beta_t}\right)}$

HOW DO WE GUARANTEE THAT H_{FINAL} GENERALIZES?

WE NEED TO MAKE SURE THAT m, # OF TRAINING POINTS IS SUFFICIENTLY LARGE.

IF γ (ACCURACY) IS INDEPENDENT FROM m, SIZE OF TRAINING SET THEN WE CAN CHOOSE m TO BE SUFFICIENTLY LARGE.

ADABOOST IS ACTUALLY A SPECIAL CASE OF AN ALGORITHM DUE TO FREUND AND SCHAPIRE CALLED "HEDGE."

HEDGE: "BEST EXPERTS" SETUP.

$c_1, \dots\dots, c_m$ AT EACH ITERATION t EXPERT c_i SUFFERS A LOSS $l_i^t \in [0,1]$

l^t A VECTOR OF LOSSES SUFFERED BY ALL EXPERTS AT t^{th} ITERATION.

INTUITION: WE WANT TO HAVE A MIXED STRATEGY OF EXPERTS
WEIGHTED AVG OF "

GOAL: SUM OF OUR LOSSES AFTER T ITERATIONS
SHOULD BE "CLOSE" TO BEST EXPERT IN HINDSIGHT.

AT EACH iteration WE MAINTAIN A SET OF WEIGHTS

$w_1, \dots, w_m$ WEIGHTED AVERAGE $= \dfrac{w_i}{\sum_i w_i} = p_i$

$w_1^t, \dots, w_m^t \rightarrow$ PROBABILITY DISTRIBUTION $\underline{p^t}$

$$p_i^t = \frac{w_i^t}{\sum_i w_i}$$

LOSS WE SUFFER AT t ITERATION IS $\underline{p^t \cdot l^t}$

WT AVERAGE OF LOSS OF EXPERTS

TOTAL LOSS WE SUFFER AFTER T ITERATIONS $= \sum_{t=1}^{T} p^t \cdot l^t \leftrightarrow$

<u>CLAIM:</u>

$$\frac{\underline{\text{YOUR LOSS}}}{T} \leq \frac{\min_i \sum_{t=1}^{T} l_i^t}{T} + \frac{O(\sqrt{T} \log n)}{T}$$

HEDGE: $\underline{w_i^{NEW}} = \underline{w_i^{OLD}} \cdot \underline{\beta}^{l_i^t}$

$w_1^t, \dots, w_m^t \rightarrow$ PROBABILITY DISTRIBUTION p^t

$$p_i^t = \frac{w_i^t}{\sum_i w_i}$$

LOSS WE SUFFER AT t ITERATION IS $p^t \cdot l^t$

WT AVERAGE OF LOSS OF EXPERTS

TOTAL LOSS WE SUFFER AFTER T ITERATIONS $= \sum_{t=1}^{T} p^t \cdot l^t$

CLAIM:

YOUR LOSS / T $\leq$

$$\frac{\min_i \sum_{t=1}^{T} l_i^t}{T} + \frac{O(\sqrt{T} \log m)}{T}$$

HEDGE: $W_i^{NEW} = W_i^{OLD} \cdot \beta^{l_i^t}$