ONLINE LEARNING:

2 Formulations

→ Mistake Bounded Model

→ Regret Minimization.

MISTAKE BOUNDED MODEL:

Cannot make more than a fixed mistake.

$$y_i = f_*(x_i)$$

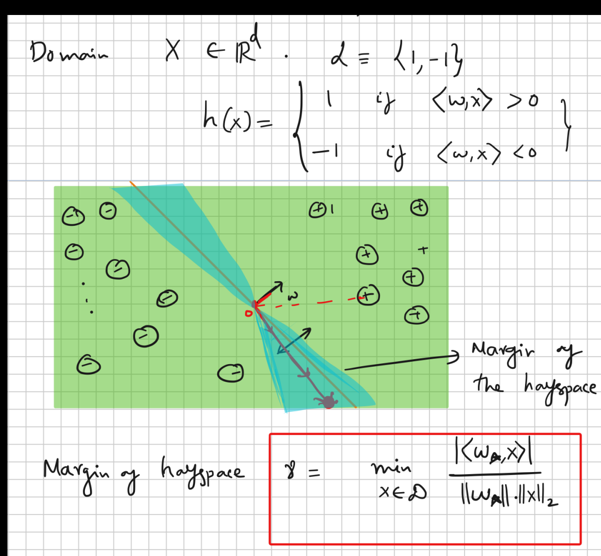
↳ ASSUMED CONSTANT!

→ Let us see if this can be applied to a very simple problem.

ONLINE LEARNING OF HALF SPACES:

$$f_*(x) = \text{sign}(\langle w_*, x \rangle) \qquad \|w_*\| = 1$$

$$\|x\| = 1$$



Given: The data is not only separated but also have a margin, γ.

$$\gamma = \min_{x \in \infty} \frac{|\langle w_*, x \rangle|}{\|w_*\| \cdot \|x\|_2} = \min_{x \in \infty} |\langle w_*, x \rangle|$$

MARGIN : $\boxed{|\langle w_*, x_i \rangle| \geq \gamma \quad \forall\, i}$

Perceptron can learn such a models for the half spaces problem.

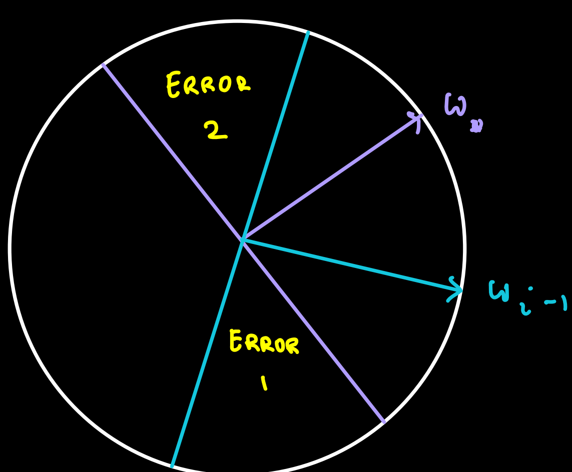
PERCEPTRON:

→ $w_0 \equiv$ random vector

→ On Day i

$$w_i = \begin{cases} w_{i-1} & \text{if no mistake} \\ w_{i-1} + y_i x_i & \text{if mistake} \end{cases}$$

Perceptron makes at most $1/\gamma^2 + 1$ mistakes ($\gamma \equiv$ margin of w_*)



$$y_i = \text{sign}(\langle w_*, x_i \rangle)$$

PROOF:

we know $\boxed{\|w_0\|^2 = \|w_*\|^2 = 1, \quad \|x_i\|^2 = 1}$

$\|w_i\|^2$ need not be 1.

On mistake: $w_i = w_{i-1} + y_i x_i$ - ①

1. $\|w_i\|^2 = \|w_{i-1} + y_i x_i\|^2$

$$= \|w_{i-1}\|^2 + \|y_i x_i\|^2 + 2\langle w_{i-1}, y_i x_i \rangle$$

$$\underbrace{y_i^2}_{1} \underbrace{\|x_i\|^2}_{1} \qquad 2 \cdot y_i \langle w_{i-1}, x_i \rangle$$

$$2 y_i \underbrace{\langle w_{i-1}, x_i \rangle}$$

$y_{pred} = \text{sign}(\langle w_{i-1}, x_i \rangle)$

mistake so $y_i \neq y_{pred}$

This is ≤ 0

So $\|w_i\|^2 \leq \|w_{i-1}\|^2 + 1$

$\|w_0\|^2 = 1$

So at any point

$\|w_n\|^2 \leq 1 + M_n \qquad \|w_n\| \leq \sqrt{1 + M_n}$

2. Dot product with w^*

$$\langle w_i, w_* \rangle = \langle w_{i-1}, w_* \rangle + \underbrace{\langle y_i x_i, w_* \rangle}$$

$y_i \langle x_i, w_* \rangle$

$y_i = \text{Sign}(\langle x_i, w_* \rangle)$

So its always positive

$|\langle x_i, w_* \rangle|$

$\geq \gamma$ [Margin]

$$\langle w_i, w_* \rangle \geq \langle w_{i-1}, w_* \rangle + \gamma$$

So if M_n mistakes

$$\langle w_n, w_* \rangle \geq \langle w_0, w_* \rangle + \gamma M_n$$

$$\|w_n\| \cdot \underbrace{\|w_*\|}_{1} \geq \langle w_0, w_* \rangle + \gamma M_n \quad \text{[Cauchy-Shwartz]}$$

$$\sqrt{1+M_n} \geq \underbrace{\langle w_0, w_* \rangle}_{\geq -1} + \gamma M_n$$

$$\sqrt{1+M_n} \geq \gamma M_n - 1$$

$$1 + M_n \geq \gamma^2 M_n^2 + 1 - 2\gamma M_n$$

$$2\gamma M_n + M_n \geq \gamma^2 M_n^2$$

$$2\gamma + 1 \geq \gamma^2 M_n$$

$$M_n \leq \frac{2}{\gamma} + \frac{1}{\gamma^2} \leq \frac{1}{\gamma^2} + 2$$

$$\gamma \leq |\langle w_*, x_i \rangle| \leq 1$$

$$M_n \leq 1/\gamma^2 + 1$$

$$\boxed{\text{PERCEPTRON AS SGD WITH HINGE LOSS!}}$$

REGRET MINIMIZATION:

LEARNING WITH EXPERTS: d: number of experts.

$$\boxed{\text{Regret}(n) = \sum_{t=1}^{n} L(t) - \min_{i=1,\dots,d} \sum_{t=1}^{n} L(i,t).}$$

* **FOLLOW THE MAJORITY:**

  ASSUMPTION: One infallible expert.

$$\boxed{\text{Regret}(T) \leq \log_2 d}$$

PROOF: $W(t)$ = # eligible experts

* $w(t) \geq 1$
* $w(t) \leq w(0) \cdot \left(\frac{1}{2}\right)^{L(t)}$

$$\leq d \cdot \left(\frac{1}{2}\right)^{L(t)}$$

$$d \cdot \left(\frac{1}{2}\right)^{L(t)} \geq 1$$

$$L(t) \leq \log_2 d \;//.$$

* WEIGHTED MAJORITY ALGORITHM:

- $w(0,i) = 1$
- Predict: Weighted majority
- If expert i made mistake:

$$w(t,i) = w(t-1,i)/2$$

$$\boxed{L(T) \leq 2.4\,(L_*(T) + \log_2 d)}$$

$L(T)$ → Total loss till T

PROOF:

$$W(t) = \sum_{i=1}^{d} w(t,i)$$ [Total weight of all experts]

If we make a mistake on day t

$$W(t) \leq \left(\frac{3}{4}\right) W(t-1)$$

If $L(t)$ = our loss after t rounds, then

$$W(t) \leq W(0) \cdot \left(\frac{3}{4}\right)^{L(t)}$$

$$\leq d \cdot \left(\frac{3}{4}\right)^{L(t)}$$

Also

$$W(t) \geq \left(\frac{1}{2}\right)^{L_*(t)}$$

Combine: $$\left(\frac{1}{2}\right)^{L_*(t)} \leq d \cdot \left(\frac{3}{4}\right)^{L(t)}$$

$$\left(\frac{4}{3}\right)^{L(t)} \leq 2^{L_*(t)} \cdot d$$

$$\log_2 4/3 \cdot L(t) \leq L_*(t) + \log_2 d$$

$$L(t) \leq 2.4\left(L_*(t) + \log_2 d\right).$$

Algorithms with $\frac{\text{Regret}(T)}{T} \to 0$ as $T \to \infty$ are

"NO-REGRET ALGORITHM".

* MULTIPLICATIVE WEIGHTS UPDATE METHOD:

- $w(0,i) = 1$
- Predict: Select expert with probability $\propto w(t-1,i)$

$$\Pr[\text{expert } i] = \frac{w(t-1,i)}{\sum_{j=1}^{d} w(t-1,j)}$$

- If expert i made mistake:

$$w(t,i) = (1-\varepsilon)\, w(t-1,i)$$

[Update all experts who made mistake]

Total loss till T.

$$\boxed{A(T) \leq (1+\varepsilon) L_*(T) + \frac{\ln d}{\varepsilon} \quad (\varepsilon < 1/2)}$$

If $\varepsilon = \sqrt{\frac{\ln d}{T}}$

$$A(T) \leq L_*(T) + 2\sqrt{T \ln d}$$

* $L(T)$ here is loss at T unlike before.

PROOF :

$$W(t) = \sum_{i=1}^{d} w(t,i)$$

$$w(T,i) = (1-\varepsilon)^{\sum_{t=1}^{T} l(t,i)}$$

i. $L(t) = E[\text{loss incurred on day } t]$ same as the loss of expert we chose

$$= \sum_{i=1}^{d} \Pr[\text{We pick expert } i] \cdot l(t,i)$$

$$= \sum_{j=1}^{d} \frac{w(t-1,i)}{\underbrace{\sum_{j=1}^{d} w(t-1,j)}_{W(t-1)}} \cdot l(t,i)$$

$$L(t) = \frac{1}{W(t-1)} \cdot \sum_{i=1}^{d} w(t-1,i) \cdot l(t,i) \quad - (1)$$

ii. $W(t) = \sum_{i=1}^{d} w(t,i)$

$$= \sum_{i=1}^{d} w(t-1,i) \cdot (1 - \varepsilon\, l(t,i))$$

$$= \sum_{i=1}^{d} w(t-1,i) - \varepsilon \sum_{i=1}^{d} w(t-1,i) \cdot l(t,i)$$

$$= W(t-1) - \varepsilon \cdot W(t-1)\, L(t) \quad \text{[From (1)]}$$

$$W(t) = W(t-1) \cdot (1 - \varepsilon L(t))$$

iii. $\boxed{1 - x \leq e^{-x} \quad \forall x}$

$$W(t) \leq W(t-1) \cdot e^{-\varepsilon L(t)}$$

$$W(T) = W(0) \cdot e^{-\varepsilon (L(1) + L(2) + \dots + L(T))}$$

$$= W(0) \cdot e^{-\varepsilon \cdot A(T)}$$

$\downarrow$ Total expected loss.

iv. $W(T) \geq (1-\varepsilon)^{L_{x}(T)}$

$$(1-\varepsilon)^{L_*(T)} \leq d \cdot e^{-\varepsilon A(T)}$$

$$L_*(T) \cdot \ln(1-\varepsilon) \leq \ln d - \varepsilon A(T)$$

$$\varepsilon A(T) \leq (-\ln(1-\varepsilon)) \cdot L_*(T) + \ln d$$

$$A(T) \leq \left(\frac{-\ln(1-\varepsilon)}{\varepsilon}\right) L_*(T) + \frac{\ln d}{\varepsilon}$$

v. $$\boxed{\frac{-\ln(1-x)}{x} \leq 1+x \quad \text{if} \quad x < 1/2}$$

$\varepsilon < 1/2$

$$A(T) \leq (1+\varepsilon) L_*(T) + \frac{\ln d}{\varepsilon}$$

BOOSTING:

ADABOOST

$\rightarrow D^{(0)} = \left(\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d}\right)$. $w(0,i) = 1$ for $i = 1, 2, \ldots, d$.

$\rightarrow h^{(0)} = WL(D^{(0)})$.

For $t = 1, \ldots, T$:

- Define $w(t,i) = \begin{cases} (1-\varepsilon)\, w(t-1,i) & \text{if "correct"} \\ w(t-1,i) & \text{if wrong.} \end{cases}$
- $D^{(t)}$ = distribution proportional to weights
- $h^t = WL(D^t)$

- Output $h = MAJ(h^0, h^1, h^2, \ldots, h^T)$.

Correct means $h^{t-1}(x^i) = y^i$

Adaboost achieves accuracy $1-\delta$ on the dataset if

$$\boxed{T \geq \frac{2\ln(1/\delta)}{(1/2 - \gamma)^2}}$$

Weak Learner: $\Pr_{(x^i, y^i)\sim D}[h(x^i) \neq y^i] \leq \gamma$

$\Rightarrow \quad T \propto \ln(1/\delta)$

WL : 60% we need 90%

$\gamma = 0.4$ $\delta = 0.1$

$$T \geq \frac{2\ln(10)}{(0.1)^2} \approx 600$$

Generate 600 weak learners using ADABOOST!

$$W(t) \leq W(t-1) \cdot (1-\varepsilon)^{L(t)}$$

$$L(i,t) = \begin{cases} 1 & \text{if } h^t(x^i) = y^i \\ 0 & \text{else} \end{cases}$$

$L(t) = E[\text{ loss we incur at time } t]$

↳ loss of the weak learner at that time

$\geq 1 - \gamma$

$$W(t) \leq W(t-1) \cdot (1-\varepsilon)^{1-\gamma}$$

$$\leq W(t-1) \cdot e^{-\varepsilon(1-\gamma)}$$

$$\boxed{W(T) \leq W(0) \cdot e^{-\varepsilon T(1-\gamma)}}$$

Basically $A(\gamma) = (1-\gamma)T$

For every bad example / expert, that was wrong for more than $T/2$ times. So weight became $(1-\varepsilon)w_{t-1}$ only less than $T/2$ times.

BAD index : $W(i,T) \geq (1-\varepsilon)^{T/2}$

$$W(T) \geq |BAD| \cdot (1-\varepsilon)^{T/2}$$

$$|BAD| \cdot (1-\varepsilon)^{T/2} \leq d \cdot e^{-\varepsilon T(1-\gamma)}$$

$$\left(\frac{|BAD|}{d}\right) \leq e^{-\varepsilon T(1-\gamma)} \cdot (1-\varepsilon)^{-T/2}$$

$$\leq e^{-\varepsilon T(1-\gamma)} \cdot e^{\varepsilon(1+\varepsilon)T/2}$$

$$\frac{-\ln(1-x)}{x} \leq 1+x \quad \text{if} \quad x < 1/2$$

$$\ln\left(\frac{1}{1-x}\right) \leq x(1+x)$$

$$\frac{1}{1-x} \leq e^{x(1+x)}$$

$$(1-x)^{-1} \leq e^{x(1+x)}$$

$$\frac{|BAD|}{d} \leq e^{-\varepsilon T((1/2-\gamma)-\varepsilon/2)}$$

Set $\varepsilon = 1/2 - \gamma$

$$\frac{|BAD|}{d} \leq e^{-T(1/2-\gamma)(1/2-\gamma)/2}$$

$$\leq e^{-T(1/2-\gamma)^2/2}$$

So if $T \geq \dfrac{2 \ln(1/\delta)}{(1/2 - \gamma)^2}$

$$\frac{|BAD|}{d} \leq \delta$$

SUMMARY:

1. Mistake Bounded : Perceptron

- Halfspaces with margin:

$$|\langle w_*, x_i \rangle| \geq \gamma \quad \forall i$$

- $$w_i = \begin{cases} w_{i-1} & \text{if no mistake} \\ w_{i-1} + y_i x_i & \text{if mistake} \end{cases}$$

$$y_i = \text{sign}(\langle w_*, x_i \rangle)$$

- Perceptron makes at most $1/\gamma^2 + 1$ mistakes

PROOF:

$$\|w_0\|^2 = \|w_*\|^2 = 1 \quad , \quad \|x_i\|^2 = 1$$

$w_i = w_{i-1} + y_i x_i$

$$\rightarrow \|w_i\|_2^2 = \|w_{i-1} + y_i x_i\|_2^2$$

$$\|w_n\|^2 \leq 1 + M_n$$

$$\rightarrow \langle w_i, w_* \rangle = \langle w_{i-1}, w_* \rangle + \langle y_i x_i, w_* \rangle$$

$$\searrow \gamma$$

$$\langle w_n, w_* \rangle = \langle w_0, w_* \rangle + \gamma M_n$$

- Perceptron <=> SGD + Hinge Loss

2. Regret Minimization

i. 1 Infallible + FTM :

$$Regret(T) \leq \log_2 d$$

ii. WMA :

$$L(T) \leq 2.4 \left(L_*(T) + \log_2 d \right)$$

iii) No - Regret : MWM:

Predict , $Pr \propto w(t-1, i)$

On Mistake : $w(t,i) = (1 - \varepsilon L(t,i)) \, w(t-1, i)$

$$A(T) \leq (1+\varepsilon) L_*(T) + \ln d / \varepsilon$$

$$\varepsilon = \sqrt{\frac{\ln d}{T}}$$

$$A(T) \leq L_*(T) + 2\sqrt{T \ln d}$$

PROOF :

$\rightarrow$ $W(t) \leq W(t-1) \cdot (1-\varepsilon)^{L(t)}$

$$W(T) \leq W(0) \cdot (1-\varepsilon)^{A(T)} \leq d \cdot e^{-\varepsilon (A(T))}$$

$\rightarrow$ $W(T) \geq (1-\varepsilon)^{L_*(T)}$

$$(1-\varepsilon)^{L_*(T)} \leq d \cdot e^{-\varepsilon (A(T))}$$

$$L_*(T)\ln(1-\varepsilon) \leq \ln d - \varepsilon A(T)$$

$$A(T) \leq \left(\frac{-\ln(1-\varepsilon)}{\varepsilon}\right) L_d(T) + \frac{\ln d}{\varepsilon}$$

3. Boosting : ADABOOST

Weak Learner : $\Pr[h(x^i) \neq y^i] \leq \gamma$

Achieves accuracy $1-\delta$ if

Number of WLs , $T \geqslant \dfrac{2\ln(1/\delta)}{(1/2 - \gamma)^2}$

PROOF :

→ $A(T) \geqslant T \times (1-\gamma)$

$$w(T) \leq w(0) \cdot (1-\varepsilon)^{A(T)} \leq d \cdot e^{-\varepsilon T(1-\gamma)}$$

→ $w(T) \geqslant |BAD| \cdot (1-\varepsilon)^{T/2}$

→ Properties:

$$1-x \leq e^{-x} \quad \forall x$$

$$\frac{-\ln(1-x)}{x} \leq 1+x \quad \text{if} \quad x < 1/2$$

QUESTIONS

$\gamma \in (0,1)$

atleast 'γd' good experts $\rightarrow$ Loss $\leq L$

$$w(T) \geq (1-\varepsilon)^L \cdot \gamma d + \text{other weights}$$

$$w(T) \leq d \cdot e^{-\varepsilon(A(T))}$$

$$(1-\varepsilon)^L \cdot \gamma \cancel{d} \leq \cancel{d} \cdot e^{-\varepsilon(A(T))}$$

$$L \ln(1-\varepsilon) + \ln\gamma \leq -\varepsilon A(T)$$

$$A(T) \geq \frac{L}{\varepsilon} \ln(1-\varepsilon) + \frac{\ln\gamma}{\varepsilon}$$