LAST TIME: Online Learning in Mistake-Bounded Model

For $t = 1, 2, \ldots$

- Receive $\underline{x}_t$
- Output $\hat{y}_t \in \{0,1\}$
- Receive $y_t \in \{0,1\}$ mistake if $\hat{y}_t \neq y_t$.

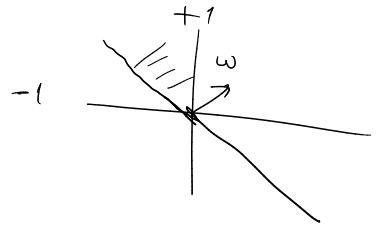
GOAL: #MISTAKES bounded even as $t \to \infty$.

(this requires that $y_t = c(x_t)$ for some $c \in C$, where C is some concept class)

PERCEPTRON: (Rosenblatt - 50s).

$f : \mathbb{R}^n \to \{-1, 1\}$

$$f_{\underline{w}}(x) = \begin{cases} +1 & \text{if } \underline{w} \cdot x \geq 0 \\ -1 & \text{if } w \cdot x < 0. \end{cases}$$



Perceptron Algorithm:

Set $\underline{w}_0 = \underline{0} \in \mathbb{R}^n$

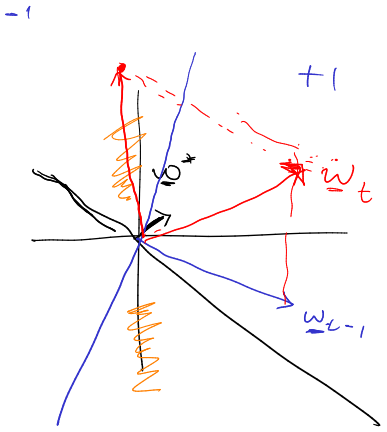
for $t = 1, 2, \ldots$ do

  receive $\underline{x}_t$

  $\hat{y}_t = \text{sign}(\underline{w}_{t-1} \cdot \underline{x}_t)$

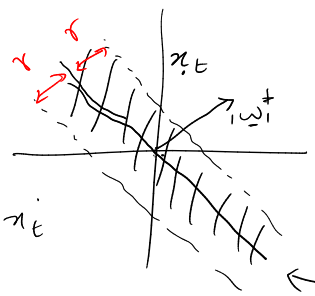
  if $\hat{y}_t \neq y_t$:

    $\underline{w}_t = \underline{w}_{t-1} + y_t \underline{x}_t$



$$\begin{cases} \text{sign}(x) = +1 & x \geq 0 \\ \phantom{\text{sign}(x)} = -1 & x < 0. \end{cases}$$

Theorem: Suppose $\|x_t\| \leq D \;\forall t$ (bounded) and $\underline{w}^*$ defines the true linear separator, i.e. $y_t = \text{sign}(\underline{w}^* \cdot \underline{x}_t) \;\forall t$. Furthermore suppose $\|\underline{w}^*\| = 1$ and that $|\underline{w}^* \cdot x_t| \geq \gamma > 0$ for all t, (margin condition) then the total #MISTAKES of perceptron $\leq O\left(D^2/\gamma^2\right)$.



← margin around the separator in which there is no data.

Lemma: Let m_t be the mistakes made up to time t

(i) $\underline{w}_t \cdot \underline{w}^* \geq \gamma \cdot m_t$

(ii) $\|\underline{w}_t\|^2 \leq m_t \cdot D^2$.

Proof By induction

(i) $\underline{w}_0 \cdot w^* = 0 \geq \gamma \cdot 0$ (base case).

if no mistake at time t, $\underline{w}_{t+1} = \underline{w}_t$ & $m_{t+1} = m_t$ so clearly holds

else $\underline{w}_{t+1} = \underline{w}_t + y_t \cdot \underline{x}_t$

$$\boxed{|z| = z\,\mathrm{sign}(z)}$$

$$\therefore \quad \underline{w}_{t+1} \cdot \underline{w}^* = \underline{w}_t \cdot \underline{w}^* + \mathrm{sign}(\underline{w}^* \cdot x_t)(\underline{w}^* \cdot x_t)$$

$$= \underline{w}_t \cdot \underline{w}^* + |w^* \cdot x_t|$$

$$\geq \underline{w}_t \cdot \underline{w}^* + \gamma \geq (m_t + 1)\gamma.$$

(ii) $\|\underline{w}_0\|^2 = 0 \leq 0 \cdot D^2$.

$\|\underline{w}_{t+1}\|^2 = \|\underline{w}_t\|^2 \leq (m_t) D^2 = (m_{t+1}) D^2$ if no mistake at time t.

else $\|\underline{w}_{t+1}\|^2 = \|\underline{w}_t\|^2 + y_t^2 \|\underline{x}_t\|^2 + \underbrace{2 y_t (\underline{w}_t \cdot \underline{x}_t)}_{\leq 0} \leq \|\underline{w}_t\|^2 + \|\underline{x}_t\|^2$

$$\boxed{\mathrm{sign}(\underline{w}_t \cdot x_t) \neq y_t}$$

$$\leq m_t D^2 + D^2$$

$$= (m_t + 1) D^2. \quad \blacksquare$$

Proof of Theorem:

$$m_t \cdot \gamma \leq w_t \cdot w^* \underset{\text{Cauchy Schwarz}}{\leq} \|\underline{w}_t\| \cdot \|w^*\| \leq \sqrt{m_t} \cdot D.$$

$$\Rightarrow \quad m_t \leq \frac{D^2}{\gamma^2} \quad \forall t.$$

HALFSPACES / LINEAR THRESHOLD FUNCTIONS

- $VC(LTF_n) = n+1$
- Linear programming (OR SVMs) to get a consistent learner.

} doesn't require a margin assumption.

MONOTONE DISJUNCTIONS:

$$f(x) = \begin{cases} 1 & \text{if } x_{i_1} \vee x_{i_2} \cdots \vee x_{i_k} \equiv \text{TRUE}. \\ 0 & \text{o/w} \end{cases}$$

$$f(x) \equiv x_1 \vee \cdots \vee x_k$$

$$f(\underline{x}) \equiv \text{sign}\Big(\underbrace{\sum_{i=1}^{k} x_i - \frac{1}{2}} \Big).$$

(Perceptron Theorem):

$$\|\tilde{w}^*\|^2 = k + 1/4$$

could scale $\tilde{\underline{w}}^*$ by $\frac{1}{\sqrt{k+\frac{1}{4}}}$ to get $\underline{w}^*$ which is a unit vector.

(i) $\|\underline{x}_t\|^2 \leq n+1$, $D \simeq \sqrt{n+1}$ $\quad \underline{x}_t \in \{0,1\}^n$. $\qquad D = \Theta(\sqrt{n})$

(ii) $\gamma \simeq \frac{1}{\sqrt{k+\frac{1}{4}}} \left(\frac{1}{2}\right) = \Theta\left(\frac{1}{\sqrt{k}}\right)$ $\qquad \gamma = \Theta\left(\frac{1}{\sqrt{k}}\right)$.

Perceptron mistake bound: $O\left(\frac{D^2}{\gamma^2}\right) = O(nk)$.

<u>GOAL</u>: Mistake bound of $O(k \log n)$.

- Sample complexity for learning disjunctions on k literals $= \text{poly}(k, \log n)$.
- Online learning $\Rightarrow$ PAC learning.

  (if we had $\Theta\left(\frac{1}{\varepsilon} \log \frac{M}{\delta}\right)$ steps without making a mistake we would find a good hypothesis

  # time steps to simulate is $O\left(\frac{M}{\varepsilon} \log \frac{M}{\delta}\right)$.

Littlestone's Algorithm / WINNOW / Weighted Majority Algorithm.

(Learning DISJUNCTIONS)

[Littleston '89 : Learning where irrelevant variables abound..]

1. $\underline{w}^0 = (1, \ldots, 1) \in \mathbb{R}^n$
2. For $t = 1, 2, \ldots$, do

   Receive $\underline{x}_t$

   $\hat{y}_t = \mathbb{1}\left(\underline{w}^{t-1} \cdot \underline{x}_t \geq \frac{n}{2}\right)$

   if MISTAKE.

   if $\hat{y}_t = 1$ and $y_t = 0$,

   set $w_i^t = 0$ for all i s.t $(x_t)_i = 1$] (A)

   else ($\hat{y}_t = 0$ & $y_t = 1$)

   set $w_i^t = 2 \cdot w_i^{t-1}$ for all i st. $(x_t)_i = 1$.] (B)

Case (A) : If i st $(x_t)_i = 1$, then "x_i" cannot be in target disjunction.

Case (B) : If weight not enough to cross majority, doubles all the weights for i st. $(x_t)_i = 1$.

Claim 1: For each i, t ; $w_i^t \leq n$

Proof : Weights increase only when $\hat{y}_t = 0$ & $y_t = 1$.

for any i st. $(x_t)_i = 1$, it must be that $(w_i^{t-1}) < n/2$.

otherwise $\underline{w}^{t-1} \cdot x_t \geq \frac{n}{2}$. and $\hat{y}_t$ would have been 1.

if you double it, we still have $w_i^t \leq n$. ∎

Claim 2: # Promotions $\equiv P$ (no. of steps of type (B))

eliminations $\equiv E$ (no. of steps of type (A))

$E \leq P + 2$.

Proof:

$$0 \leq \sum_i w_i^t \leq \underbrace{n}_{\text{at time } 0} + \underbrace{P \cdot \frac{n}{2}}_{\text{every promotion increase total wt by at most } n/2} - \underbrace{E \cdot n/2}_{\text{every elimination step decreases total wt by at least } n/2}$$

∎

Claim 3: $P \leq k \log_2 n \quad \forall t$.

Proof : Each promotion step doubles the weight of at least one variable in the target disjunction.

Mistakes $= P + E \leq 2P + 2 \leq 2k \log_2 n + 2$. ∎