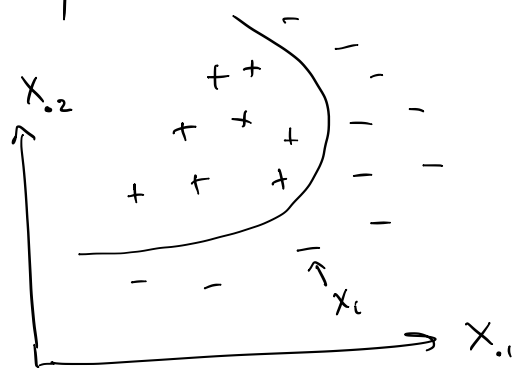
Separation happens - points can often be separated by a hyperplane in high dimensions.

If you want curvy decision boundaries, use a trick - map to a higher dim space



$$\vec{x}_i = [x_{i1}, x_{i2}] \longrightarrow [x_{i1}, x_{i2}, x_{i1}^5] = x_i^{new}$$

$$f(\vec{x}_i) = \vec{\omega} \cdot \vec{x}_i \longrightarrow f^{new}(\vec{x}_i) = \vec{\omega}^{new} \cdot x_i^{new}$$

$$= \omega_1 x_{i1} + \omega_2 x_{i2} \qquad = \omega_1^{new} x_{i1} + \omega_2^{new} x_{i2} + \omega_3^{new} x_{i1}^5$$

Intercepts are easy

$\vec{x}_i = [x_{i1}, x_{i2}] \longrightarrow \tilde{x}_i = [x_{i1}, x_{i2}, 1]$

$f(\vec{x}_i) = \omega \cdot x + b$

$= \omega_1 x_{i1} + \omega_2 x_{i2} + b$

$\tilde{f}(x_i) = \tilde{\omega} \cdot \tilde{x}_i$

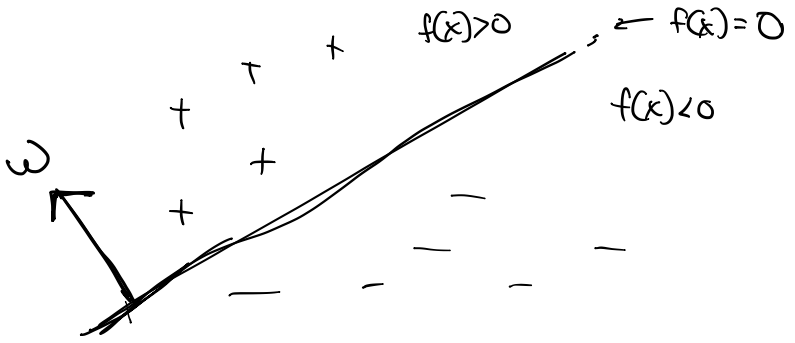
$= \tilde{\omega}_1 x_{i1} + \tilde{\omega}_2 x_{i2} + \underbrace{\tilde{\omega}_3 \cdot 1}_{b}$

$= \omega \cdot x + b$

A hyperplane is represented by a vector that is perpendicular to it.

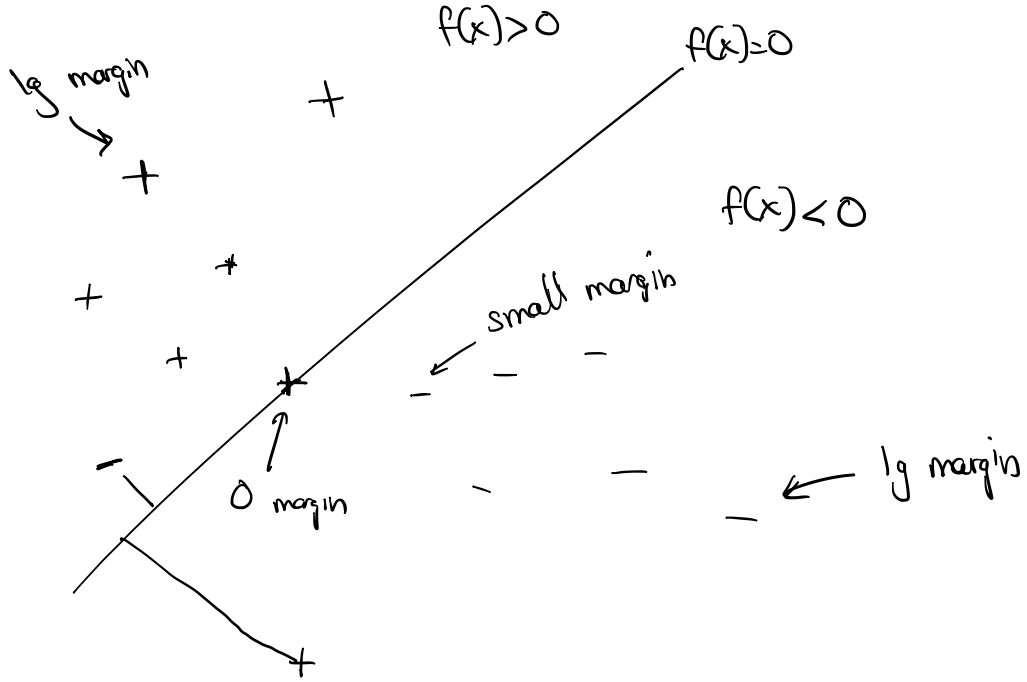
$$\hat{y} = \text{sign}(f(x)) = \text{sign}(\vec{w} \cdot \vec{x})$$

f is hyperplane



decision bdry is $\{x : \vec{w} \cdot \vec{x} + b = 0\}$

Margins - how far away we are from the decision boundary in the correct direction. margin = $y \cdot f(x) = y \cdot (\vec{w} \cdot \vec{x} + b)$

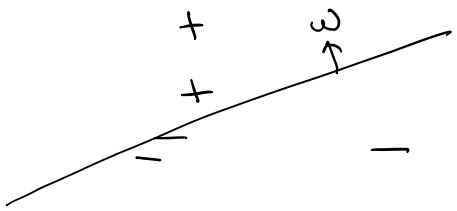


Perceptron idea

data: $(\vec{x}_1, y_1), \ldots (\vec{x}_n, y_n)$

if point is correctly classified, does nothing

if point is incorrectly classified, move decision boundary towards it



$\omega^{(t)}$ ~ model at time t

if i is correct $\quad \text{sign}(\vec{\omega}^{(t)} \cdot x_i) \cdot y_i = 1$

if i is misclassified $\quad \text{sign}(\vec{\omega}^{(t)} \cdot x_i) \cdot y_i = -1$

and $\quad y_i (\vec{\omega}^{(t)} \cdot \vec{x}_i) < 0$

want $y_i (\vec{\omega}^{(t)} \cdot \vec{x}_i) > 0$ for all i

if i is misclassified at t

$$\vec{\omega}^{(t+1)} = \vec{\omega}^{(t)} + y_i x_i$$

helpful because

$$\underbrace{y_i (\vec{\omega}^{(t+1)} \cdot \vec{x}_i)}_{\text{want this large}} = y_i \left((\vec{\omega}^{(t)} + y_i \vec{x}_i) \cdot x_i \right) = y_i (\vec{\omega}^{(t)} \cdot \vec{x}_i) + y_i (y_i \vec{x}_i \cdot \vec{x}_i)$$

$$= \underbrace{y_i (\vec{\omega}^{(t)} \cdot \vec{x}_i)}_{\text{negative}} + \underbrace{\|\vec{x}_i\|^2}_{\text{positive}}$$

Perceptron Algorithm

input : $\{(\vec{x}_i, y_i)\}_{i=1}^{n}$

initialize $\vec{w}^{(1)} = (0, 0, \cdots, 0)$

for $t = 1, 2, \cdots$

locate an i such that $y_i(\vec{w}^{(t)} \cdot \vec{x}_i) \leq 0$.

if none, stop and output $w^{(t)}$

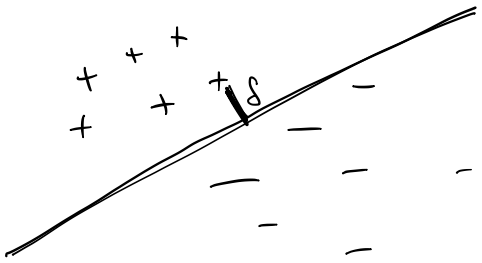
else $\vec{w}^{(t+1)} = \vec{w}^{(t)} + y_i x_i$

end

end

Perceptron Convergence Bound

Assume $\{(\vec{x}_i, y_i)\}_{i=1}^{n}$ are separable by linear functions, so there is a good separator $\vec{\omega}$ s.t. $\|\vec{\omega}\|_2 = 1$ and $y_i(\vec{\omega} \cdot \vec{x}_i) \geq \delta$ for all i. Assume $\max_i \|x_i\|_2 = 1$ (normalize)



Theorem: The perceptron algorithm makes at most $1/\delta^2$ mistakes

Define ω^* to be a "good separator" $y_i(\vec{\omega}^* \cdot \vec{x}_i) \geq \delta \ \forall i$ and $\|\vec{\omega}^*\| = 1$

at T iterations, show $\vec{\omega}^{(T+1)}$ is "close to" ω^*. $\Rightarrow T \leq \frac{1}{\delta^2}$

show $\delta\sqrt{T} \overset{?}{\leq} \dfrac{\vec{\omega}^* \cdot \vec{\omega}^{(T+1)}}{\|\vec{\omega}^*\| \|\vec{\omega}^{(T+1)}\|} \overset{\cos \leq 1}{\leq} 1.$

$$\delta\sqrt{T} \leq 1$$

$$\sqrt{T} \leq 1/\delta \quad \text{so} \quad T \leq \frac{1}{\delta^2}$$

Step 1:

$$(\vec{w}^* \cdot \vec{w}^{(t+1)}) - (\vec{w}^* \cdot \vec{w}^{(t)}) = \vec{w}^* (w^{(t+1)} - w^{(t)}) = \vec{w}^* \cdot y_i \vec{x}_i$$

(mistake at t)

$$= y_i (\vec{w}^* \cdot \vec{x}_i) \geq \delta$$

$$(w^* \cdot w^{(T+1)}) = (\vec{w}^* \cdot \vec{w}^{(T+1)}) - (\vec{w}^* \cdot w^{(1)})$$

($w^{(1)} = (0,0,0,\ldots)$)

$$= \sum_{t=1}^{T} (\vec{w}^* \cdot \vec{w}^{(t+1)}) - (\vec{w}^* \cdot \vec{w}^{(t)}) \geq T\delta$$

Step 2: for each t

$$\|\vec{w}^{(t+1)}\|^2 = \|\vec{w}^{(t)} + y_i \vec{x}_i\|^2 = \|\vec{w}^{(t)}\|^2 + 2 \underbrace{y_i (\vec{w}^{(t)} \cdot \vec{x}_i)}_{\leq 0} + y_i^2 \underbrace{\|\vec{x}_i\|^2}_{1}$$

($y_i^2 = 1$)

$$\|\vec{w}^{(t+1)}\|^2 \leq \|\vec{w}^{(t)}\|^2 + 1$$

so $\|\vec{w}^{(T+1)}\|^2 \leq T$ ($\|\vec{w}^{(1)}\| = 0$) $\Rightarrow$ $\boxed{\|\vec{w}^{(T+1)}\| \leq \sqrt{T}}$.

Step 3: combine

$$1 \geq \frac{\vec{w}^* \cdot \vec{w}^{(T+1)}}{\underbrace{\|\vec{w}^*\|}_{1} \|\vec{w}^{(T+1)}\|} \geq \frac{T\delta}{\sqrt{T}} = \sqrt{T}\,\delta$$

$\delta\sqrt{T} \leq 1$

$\sqrt{T} \leq 1/\delta$

$T \leq 1/\delta^2$ ∎

Winnow Algorithm (Littlestone, 1988)

Input: $\{(\vec{x}_i, y_i)\}_{i=1}^{n}$, parameter $\eta > 0$

Initialize: $w_j^{(0)} = \frac{1}{p}$ $\quad j = 1 \text{ -- } p$ (p features)

For $t = 0, 1, 2, \ldots$

locate i s.t. $y_i (w^{(t)} \cdot x_i) \leq 0$.

if none, stop and output $w^{(t)}$

else $\forall j \quad w_j^{(t+1)} = \frac{w_j^{(t)} e^{\eta y_i x_{ij}}}{Z_t}$

end

Z_t ← normalization

$$Z_t = \sum_j w_j^{(t)} e^{\eta y_i x_{ij}}$$

idea:

reward features that agree with label:

if $\underbrace{\text{sign}(x_{ij})}_{} = y_i$ then $w_j^{(t+1)} \propto w_j^{(t)} \cdot e^{\eta(+ve)}$ ←

punish features that disagree

if $y_i \neq \text{sign}(x_{ij})$ then $w_j^{(t+1)} \propto w_j^{(t)} e^{\eta(-ve)}$ ←

Winnow Algorithm (Littlestone 1988)

Input : $\{(\bar{x}_i, y_i)\}_{i=1}^{n}$, parameter $\eta > 0$

Initialize $w_j^{(0)} = \frac{1}{p}$ $j = 1 \dots p$ (p features)

For $t = 0, 1, 2 \dots$

locate i s.t. $y_i(w^{(t)} \cdot x_i) \leq 0$.

if none, stop and output $w^{(t)}$

else $\forall j$ $w_j^{(t+1)} = \dfrac{w_j^{(t)} \cdot e^{\eta y_i x_{ij}}}{Z_t}$

Z_t: normalization

$$Z_t = \sum_j w_j^{(t)} e^{\eta y_i x_{ij}}$$

End

idea:

reward features that agree with the label

→ if $\underbrace{\text{sign}(x_{ij}) = y_i}$ then $w_j^{(t+1)} \propto w_j^{(t)} \cdot e^{\eta(+ve)}$

→ if $\text{sign}(x_{ij}) \neq y_i$ then $w_j^{(t+1)} \propto w_j^{(t)} \cdot e^{\eta(-ve)}$

punish features that disagree with y_i.

Winnow Convergence Bound

Assume $\max_i \|x_i\|_\infty = 1$ (normalize)

Assume there is w^* s.t. $\underbrace{y_i(\vec{w}^* \cdot \vec{x}_i) \geq \delta}_{\text{separates}}$ $\forall i$

and $w^*_j \geq 0$ $\forall j$ and $\|w^*\|_1 = 1$

Theorem The Winnow Alg makes at most

$$T \leq \frac{\ln p}{n\delta + \ln\left(\frac{2}{e^{\eta} + e^{-\eta}}\right)} \text{ mistakes.}$$

proof idea: Show $\vec{w}^{(t)}$ is closer to $\vec{w}^*$ at each iteration in terms of KL divergence.

$$KL(\vec{a} \| \vec{b}) = \sum_j a_j \log\left(\frac{a_j}{b_j}\right)$$

Why is it a "distance"? If $\vec{a} = \vec{b}$, $\log\left(\frac{a_j}{b_j}\right) = 0$

so $KL(\vec{a}, \vec{b}) = 0$

Turns out $KL(\vec{a}, \vec{b}) \geq 0$

$$\Phi_t = KL(\omega^* \| \omega^{(t)}) = \sum_j \omega_j^* \ln\left(\frac{\omega_j^*}{\omega_j^{(t)}}\right)$$

$$\Phi_t - \Phi_{t+1} = \sum_j \omega_j^* \ln\left(\frac{\omega_j^*}{\omega_j^{(t)}}\right) - \sum_j \omega_j^* \ln\left(\frac{\omega_j^*}{\omega_j^{(t+1)}}\right)$$

$$= \sum_j \omega_j^* \ln\left(\frac{\omega_j^{(t+1)}}{\omega_j^{(t)}}\right)$$

$$\overset{alg}{=} \sum_j \omega_j^* \ln\left(\frac{e^{\eta y_i x_{ij}}}{Z_t}\right)$$

$$= \sum_j \omega_j^* \ln\left(e^{\eta y_i x_{ij}}\right) - \sum_j \omega_j^* \ln Z_t$$

$$= \sum_j \omega_j^* \eta y_i x_{ij} - \ln Z_t \underbrace{\sum_j \omega_j^*}_{1}$$

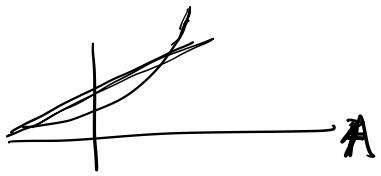
$$= \eta y_i \underbrace{\vec{\omega}^* \cdot \vec{x}_i}_{\geq \delta \text{ by def } \omega^*} - \ln Z_t$$

$$\geq \eta\delta - \ln(Z_t)$$

$$Z_t = \sum_j \underbrace{w_j^{(t)} \cdot e^{\eta y_i x_{ij}}}$$

Use a bound of an exponential by a linear function

$$e^{\eta A} \leq \left(\frac{1+A}{2}\right) e^{\eta} + \left(\frac{1-A}{2}\right) e^{-\eta} \quad \text{holds} \quad -1 \leq A \leq 1$$



$$Z_t \leq \sum_j w_j^{(t)} \left(\frac{1 + y_i x_{ij}}{2}\right) e^{\eta} + w_j^{(t)} \left(\frac{1 - y_i x_{ij}}{2}\right) e^{-\eta}$$

$A = y_i x_{ij}$

$$= \frac{e^{\eta} + e^{-\eta}}{2} \underbrace{\sum_j w_j^{(t)}}_{1} + \frac{e^{\eta} - e^{-\eta}}{2} \underbrace{\sum_j w_j^{(t)} y_i x_{ij}}_{y_i (\vec{w}^{(t)} \cdot \vec{x}_i)}$$

1 ← i is misclassified at t
0

$$\leq \frac{e^{\eta} + e^{-\eta}}{2}$$

$$-\ln(Z_t) \geq -\ln\left(\frac{e^{\eta} + e^{-\eta}}{2}\right)$$

Winnow Convergence Bound

Assume $\max_i \|x_i\|_\infty = 1$ (normalize)

Assume there is w^* s.t. $\underbrace{y_i(\vec{w}^* \cdot \vec{x}_i) \geq \delta}_{\text{separates}}$ $\forall i$ ↑

and $w^*_j \geq 0$ $\forall j$ and $\|w^*\|_1 = 1$

Theorem the Winnow Alg makes at most

$$T \leq \frac{\ln p}{\eta\delta + \ln\left(\frac{2}{e^{\eta} + e^{-\eta}}\right)}$$ ← mistakes.

So far: $\Phi_t - \Phi_{t+1} \geq \eta\delta - \ln(Z_t) \geq \eta\delta - \ln\left(\frac{e^{\eta} + e^{-\eta}}{2}\right)$

↑ ↑ ↑ ↑

Λ !!

$$\Phi_0 - \Phi_T = \sum_{t=0}^{T-1} (\Phi_t - \Phi_{t+1}) \geq T\Lambda \quad \text{lower bound}$$

$$\Phi - \Phi_T \leq \Phi_0 = KL(\vec{w}^* \| \vec{w}_0) = \sum_j w^*_j \ln\left(\frac{w^*_j}{1/p}\right)$$

↑ KL divergence is always non negative

$\frac{1}{p}$ ↑ KL ↑

$$= \sum_j w^*_j \ln(p\, w^*_j) \leq \sum_j w^*_j \underbrace{\ln(p\,1)}_{\leq \ln p} \leq \ln p \underbrace{\sum_j w^*_j}_{1}$$

↑ $0 \leq w^*_j \leq 1$

$\ln p$ upper bd

So $T\Lambda \leq \ln p \Rightarrow T \leq \frac{\ln p}{\Lambda}$. ∎

Winnow Convergence Bound

Assume $\max_i \|x_i\|_\infty = 1$ (normalize)

Assume there is w^* s.t. $\underbrace{y_i(\vec{w}^* \cdot \vec{x}_i) \geq \delta}_{\text{separates}}$ $\forall i$

and $w_j^* \geq 0$ $\forall j$ and $\|w^*\|_1 = 1$

Theorem The Winnow Alg makes at most

$$T \leq \frac{\ln p}{n\delta + \ln\left(\frac{2}{e^{\eta} + e^{-\eta}}\right)} \text{ mistakes.}$$

What value for η? Set it to minimize bound

$$\text{bound}(\eta) = \frac{\ln p}{\Lambda(\eta)} \qquad \frac{d(\text{bound}(\eta))}{d\eta} = 0$$

$$\Downarrow$$

$$\eta = \frac{1}{2}\ln\left(\frac{1+\delta}{1-\delta}\right)$$

Corollary: For Winnow if $\eta = \frac{1}{2}\ln\left(\frac{1+\delta}{1-\delta}\right)$ then

$$\#\text{mistakes} = T \leq \frac{\ln p}{n\delta + \ln\left(\frac{2}{e^{\eta} + e^{-\eta}}\right)} = \frac{2\ln p}{\delta^2}.$$

Corollary: For Winnow if $\eta = \frac{1}{2}\ln\left(\frac{1+\delta}{1-\delta}\right)$ then

$$\#\text{mistakes} = T \le \frac{\ln p}{\eta\delta + \ln\left(\frac{2}{e^{\eta}+e^{-\eta}}\right)} = \frac{2\ln p}{\delta^2}.$$

how to compare with perceptron?

perceptron

$$T \le \frac{1}{\delta^2}$$

winnow

$$T \le \frac{2\ln p}{\delta^2}$$

not a big deal

very similar

additive

$$\vec{w}^{(t+1)} = \vec{w}^{(t)} + \cdots$$

$\|\vec{x}_i\|_2 \le 1 \quad \forall i$

$\|\vec{w}^*\|_2 = 1$

SVM-like

multiplicative

$$\vec{w}^{(t+1)} = \vec{w}^{(t)} \cdot$$

$\|x_i\|_\infty \le 1 \quad \forall i$

$\|\vec{w}^*\|_1 \le 1$

AdaBoost