CS 5/4780
Perceptron

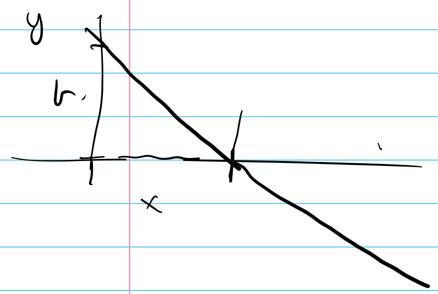
Rosenblatt '57

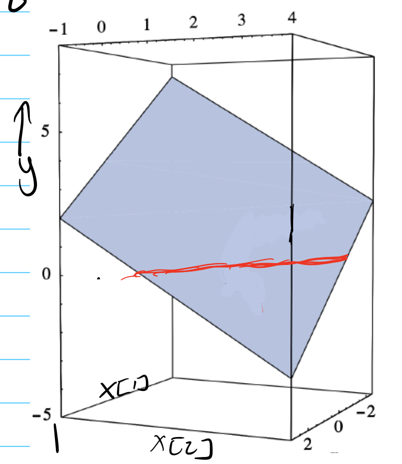
Assumption: the classes are "linearly separable"
(All positive examples are on one side of a plane and negative examples on the other side)

Binary Labels : $C = \{+1, -1\}$

Inner product: $w, x \in \mathbb{R}^d$ $\quad w^T x = \sum_{i=1}^{d} w[i] \cdot x[i]$

Add in bias: $y = w^T x + b$



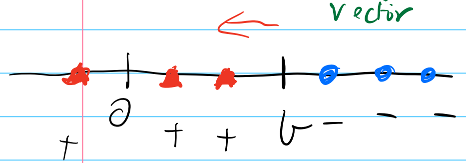


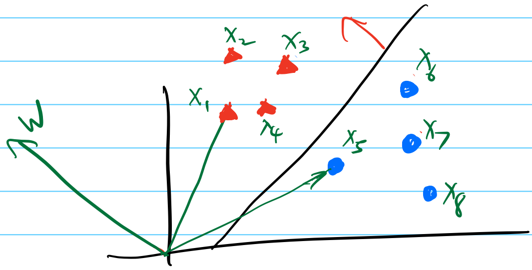
"Half space"

$h_{w,b}(x) = \text{sign}(w^T x + b)$

weight vector

bias





$w^T x = \|w\| \, \|x\| \cos_{w,x}$

Absorbing bias in $d+1$ dimensions

By increasing dimension of features by 1 more, can you find a way to encode bias in just $\text{sign}(w^T x)$?

$$x \to \begin{bmatrix} x \\ 1 \end{bmatrix} \qquad w, b \to \begin{bmatrix} w \\ b \end{bmatrix}$$

$$\begin{bmatrix} w \\ b \end{bmatrix}^T \begin{bmatrix} x \\ 1 \end{bmatrix} = w^T x + b$$

After concatenating 1 to x's, use $h_w(x) = \text{sign}(w^T x)$

Observation: if $h_w(x)$ makes a mistake on (x, y), then, $y \cdot w^T x < 0$

Perceptron Algorithm:

```
Initialize  w = 0
While TRUE:
    m = 0                         # set no. of mistakes to 0
    For i = 1 to n:
        if  y_i . w^T x_i <= 0    # is mistake
            w <- w + y_i x_i      # update
            m <- m+1              # increment m
        endif
    END For
    if ( m==0 )
        BREAK;
    end if
END WHILE
```

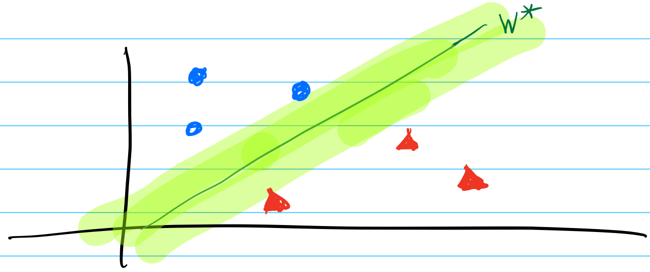
Update $w \leftarrow w + y_i x_i$ improves on (x_i, y_i)

$$y_i (w + y_i x_i)^T x_i = y_i w^T x_i + x_i^T x_i$$

improvement

When will perceptron Algo. Converge?

When data is linearly separable.



$\exists\, w^*$ s.t.

$\forall i\ y_i x_i^T w^* > 0$

$\|w^*\| = 1$

Margin: $\gamma = \min_{i \in [n]} y_i x_i^T w^* > 0$ Assume $\forall i,\ \|x_i\| \le 1$

How much time does it take to converge?

Thm: Perceptron makes at most $1/\gamma^2$ updates.

Intuition: After each mistake we update as

$w \leftarrow w + y_i x_i$

1. After each update, $\|w\|^2$ increases by at most 1

Hence M updates implies $\|w\|^2 \le M$ and $\|w\| \le \sqrt{M}$

2. After each update, $w^T w^*$ improves by atleast γ

in at most $\frac{1}{\gamma^2}$ updates

Hence $M\gamma \le w^T w^* \le \|w\| \le \sqrt{M}$ and so $M \le \frac{1}{\gamma^2}$