Perceptron

Binary Classification

$(x_1, y_1), (x_2, y_2), \ldots$

feature vec $x_i \in \mathbb{R}^d$ Label $y_i \in \{\pm 1\}$

$h(x) = y$

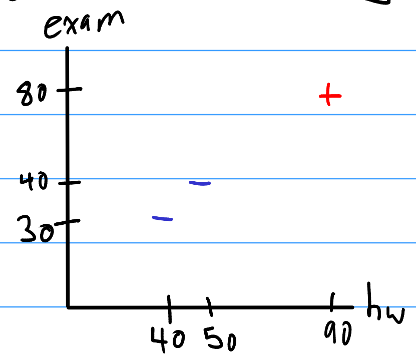
$x_i = (hw, exam)$

$y_i =$ Passed?

$x_1 = (90, 80),\ y_1 = 1$

$x_2 = (40, 30),\ y_2 = -1$

$x_3 = (50, 40),\ y_3 = -1$



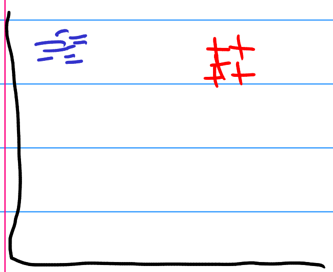
$x = (50, 60)$ $y = ?$

Statistical Learning

$(x_1, y_1), \ldots, (x_n, y_n) \overset{i.i.d.}{\sim} P$

Goal: Learn $h(\cdot): \mathbb{R}^d \to \{\pm 1\}$ s.t. $\Pr_{(x,y)\sim P}[h(x) = y]$ is large

Bayes classifier



Online Learning.

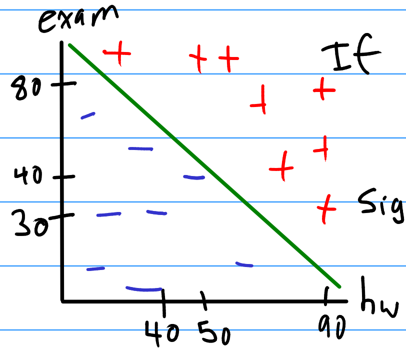
At each time $i = 1, 2, \ldots,$

1. Receive x_i,
2. Choose h_i, predict $\hat{y}_i = h_i(x_i)$
3. View y_i. Suffer mistake if $y_i \neq \hat{y}_i$

Perceptron

If average of hw + exams > 0.5
then pass

$x_1=(90,80)$, $y_1=1$
$x_2=(40,30)$, $y_2=-1$
$x_3=(50,40)$, $y_3=-1$



If $0.5\cdot hw + 0.5\cdot exam - 0.5 > 0$,
then pass

$\text{sign}(\langle(0.5,0.5),(hw,exam)\rangle - 0.5)$

$x_i=(hw,exam)$

$y_i = \text{sign}(\langle(0.5,0.5), x_i\rangle - 0.5)$

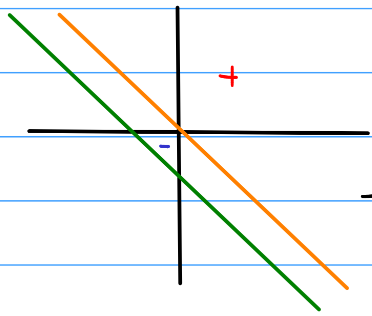
Algorithm: The Perceptron (Rosenblatt 1958)

Input: Dataset $\mathcal{D} = \{\!\{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\} : i = 1, \ldots, \mathsf{n}\}\!\}$, initialization $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$, threshold $\delta \geq 0$

Output: approximate solution $\mathbf{w}$ and b

```
for k = 1, 2, ... do
    receive training example index I_k ∈ {1, ..., n}          // the index I_k can be random
    if y_{I_k}(w^T x_{I_k} + b) ≤ δ then
        w ← w + y_{I_k} x_{I_k}                                // update only after making a "mistake"
        b ← b + y_{I_k}
```

$x_1 (1,1)$ $y_1=1$
$x_2=(-\frac{1}{4},-\frac{1}{4})$, $y_2=-1$



$w=\vec{0}$, $b=0$. $\delta=0$

$w \leftarrow w + (1,1) = (1,1)$

$b \leftarrow b + 1 = 1$

$-1(-\frac{1}{2}+1) = -\frac{1}{2} < 0$

$w \leftarrow w + (-\frac{1}{4},-\frac{1}{4})(-1) = (\frac{5}{4},\frac{5}{4})$

$b \leftarrow b + (-1) = 1 - 1 = 0$

Padding + Pre-multiplication

$y_i = \text{sign}(\langle w, x_i \rangle + b) \quad \forall i \in [n]$

$\quad = \text{Sign}(\langle (w \ \ b), (x_i \ \ 1) \rangle)$

$\quad \Leftrightarrow y_i \langle (w \ \ b), (x_i \ \ 1) \rangle > 0$

$(x_i, y_i) \Rightarrow a_i \triangleq y_i (x_i \ \ 1)$

$A = \begin{bmatrix} a_1 \\ \cdots \\ a_n \end{bmatrix} \in \mathbb{R}^{n \times (d+1)}$

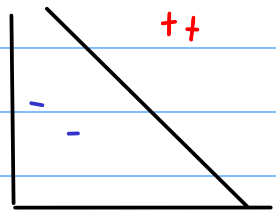
$\langle a_i, w \rangle > 0 \quad \forall i$

$Aw > \vec{0}$ (entrywise)

Linear Separability

$\exists w \text{ st } \langle a_i, w \rangle \geq s > 0 \quad \forall i$

$\Leftrightarrow \quad Aw \geq s\vec{1}$, where $s > 0$



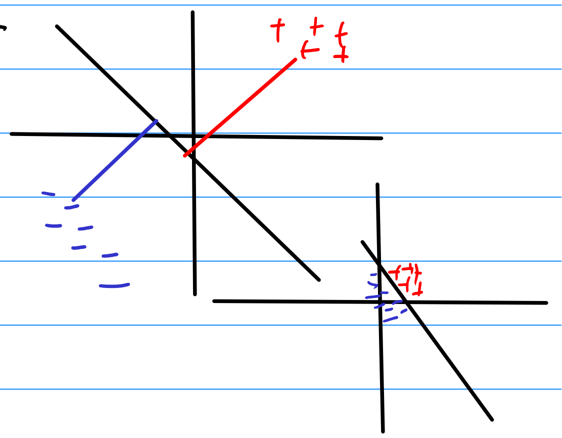
Error Bound + Margin

Suppose $\exists w, s > 0$ s.t. $Aw \geq s\mathbb{1}$.

Then perceptron converges in $\leq R^2 \cdot \left(\frac{\|w\|_2^2}{s^2}\right)$ steps, $R = \max_i \|a_i\|_2$.

$$\min_{(w,s):\, Aw \geq s\mathbb{1}} \frac{\|w\|_2^2}{s^2} = \min_{(w,s):\, \|w\|_2 = 1,\, Aw \geq s\mathbb{1}} \frac{1}{s^2}$$

$$= \frac{1}{\left(\max_{(w,s):\|w\|_2=1,\, Aw \geq s\mathbb{1}} s\right)^2} = \left(\frac{1}{\underbrace{\max_{\|w\|_2=1} \min_i \langle a_i, w \rangle}_{\text{margin } \gamma}}\right)^2$$



Uniqueness?

No. SVM

Non-Separable?
- Perceptron cycles
- Not the right algo.

When to stop?
- all pts are correct
- error stops decreasing
- weights converge
- fixed # of iters

($k=2 \rightarrow$ binary)

Multiclass
- one-vs-all - Train k # of classes classifiers Pick largest $\langle w_i, x \rangle + b$ i
- one-vs-one - k^2