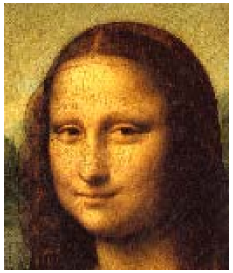
Linear Classification

6.036 Introduction to Machine Learning

Classification

Is this a face image?



y= 1

y= −1

Labels/outputs: $\mathbf{y} \in \{-1,1\}$

Feature vectors/input: $x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \in \mathrm{R}^d$

Training set: $\mathrm{S_n} = \{(x^{(i)}, y^{(i)}), i = 1 \ldots n\}$

classifier/hypothesis/separator

h: $\mathrm{R}^{d} \rightarrow \{-1,1\}$

set of classifiers/hypothesis class

h $\in \mathrm{H}$

training error

$$\mathcal{E}_{\mathrm{n}}(\mathrm{h}) = \frac{1}{n} \sum_{i=1}^{n} [[h(x^{(i)}) \neq y^{(i)}]]$$

test/generalization error

$$\mathcal{E}(\mathrm{h}) = \frac{1}{n'} \sum_{i=n+1}^{n+n'} [[h(x^{(i)}) \neq y^{(i)}]], \ (n' \rightarrow \infty)$$

Set of linear classifiers

$$h(x;\theta,\theta_0)=\operatorname{sign}(\theta\cdot x+\theta_0)=\begin{cases}+1, & \theta\cdot x+\theta_0>0\\ -1, & \theta\cdot x+\theta_0\le 0\end{cases}$$

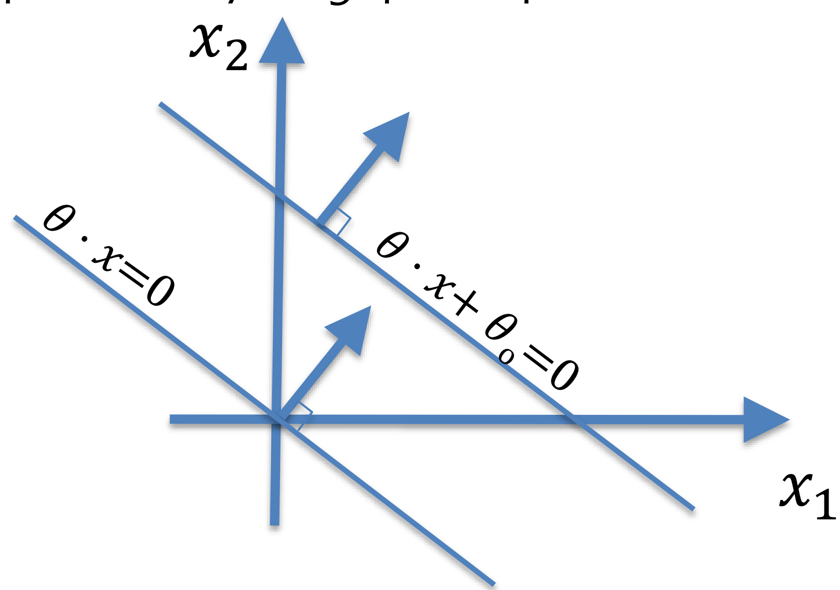
$\theta \in \mathrm{R}^d$

$\theta_0 \in \mathrm{R}$

$$\begin{bmatrix}\theta_1\\ \vdots\\ \theta_d\end{bmatrix}\cdot\begin{bmatrix}x_1\\ \vdots\\ x_d\end{bmatrix}$$

decision boundary/plane/hyperplane

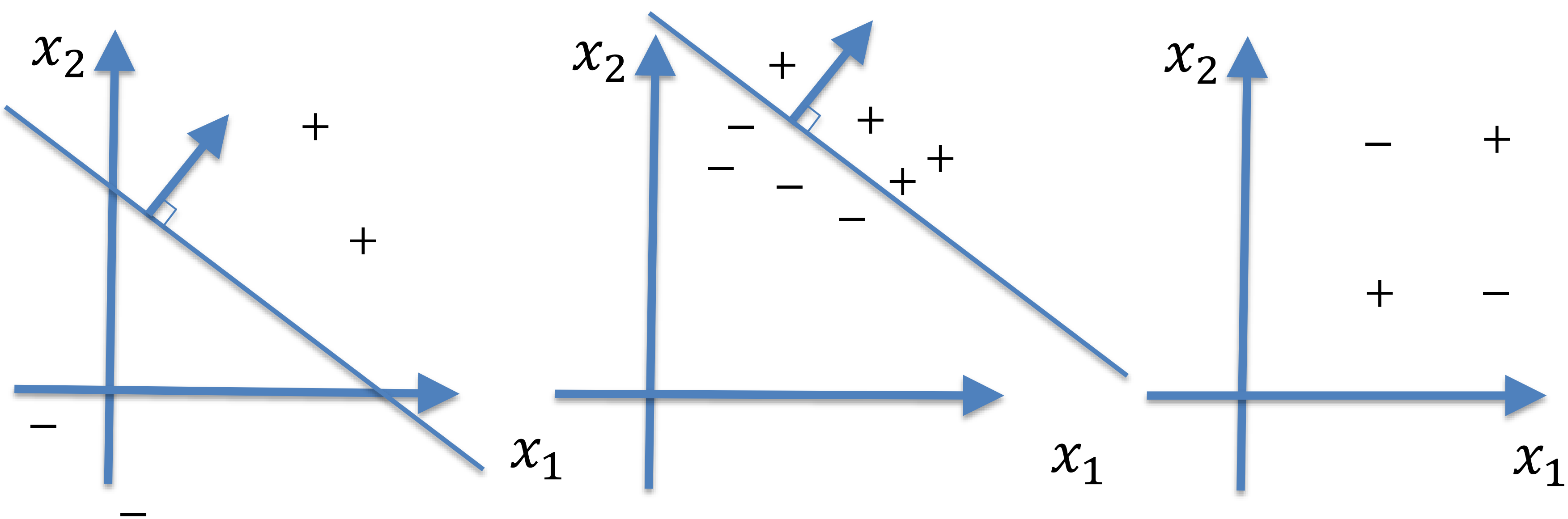
$$\{x:\theta\cdot x+\theta_0=0\}$$



Linear separation

Training examples $S_n = \{(x^{(i)}, y^{(i)}), i = 1 \ldots n\}$ are linearly seaparable if $\exists \hat{\theta}, \hat{\theta}_0$

s.t. $y^{(i)}(\hat{\theta} \cdot x^{(i)} + \hat{\theta}_0) > 0$, for all $i = 1 \ldots n$



Perceptron algorithm

Perceptron ($\left\{\left(x^{(i)}, y^{(i)}\right), i = 1, \ldots, n\right\}, T$)

$\theta^{(0)} := 0; \theta_0^{(0)} := 0;\ k := 0$

for $t = 1, \ldots \mathrm{T},\ \mathrm{do}$

for $i = 1, \ldots n,\ \mathrm{do}$

if $y^{(i)}(\theta^{(k)} \cdot x^{(i)} + \theta_0^{(k)}) \leq 0$ then //mistake

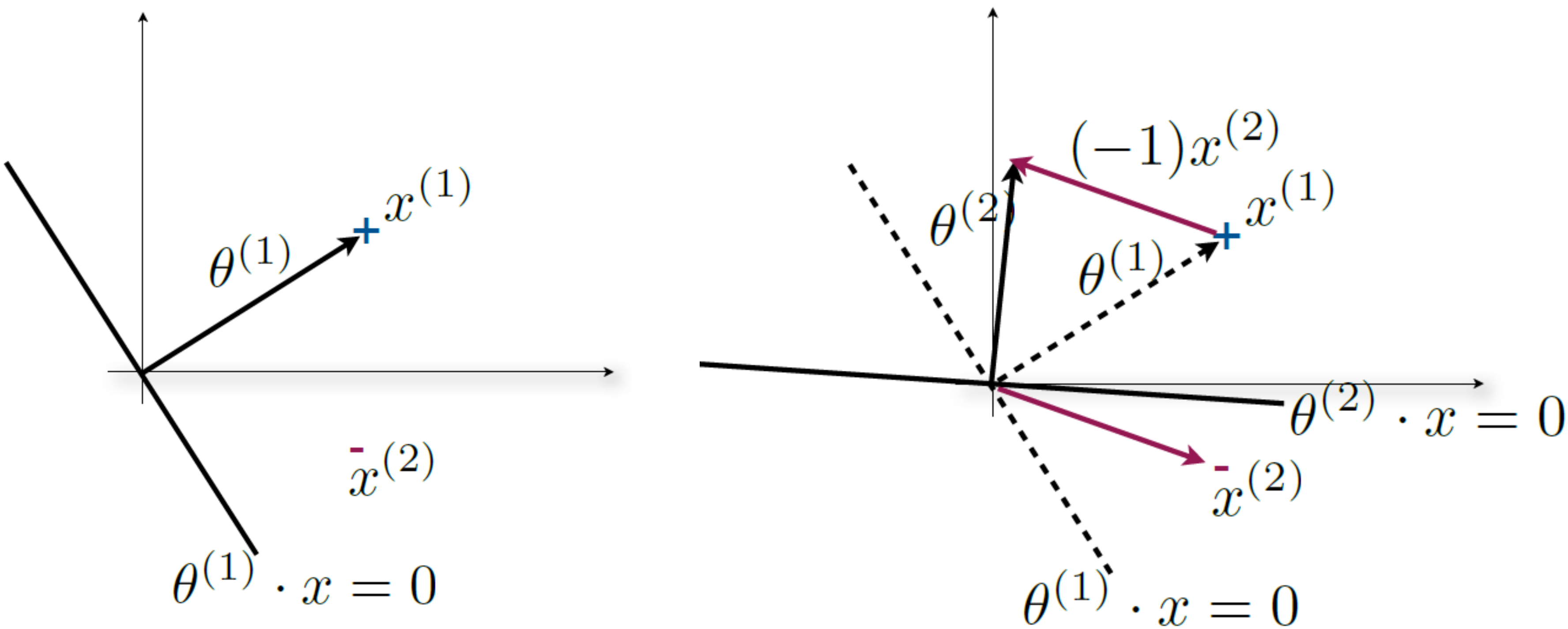
$k := k + 1$

$\theta^{(k)} := \theta^{(k-1)} + y^{(i)} x^{(i)}$

$\theta_0^{(k)} := \theta_0^{(k-1)} + y^{(i)}$

return $\theta^{(k)}$, $\theta_0^{(k)}$

Update example (no offset)



$$\theta^{(k)} := \theta^{(k-1)} + y^{(i)} x^{(i)}$$

Perceptron algorithm

Perceptron ($\{(x^{(i)}, y^{(i)}), i = 1, \ldots, n\}, T$)

$\theta^{(0)} := 0, \theta_0^{(0)} := 0, k = 0$

for $t = 1, \ldots \mathrm{T}$, do

for $i = 1, \ldots n$, do

if $y^{(i)}(\theta^{(k)} \cdot x^{(i)} + \theta_0^{(k)}) \leq 0$ then

$k := k + 1$

$\theta^{(k)} := \theta^{(k-1)} + y^{(i)} x^{(i)}$

$\theta_0^{(k)} := \theta_0^{(k-1)} + y^{(i)}$

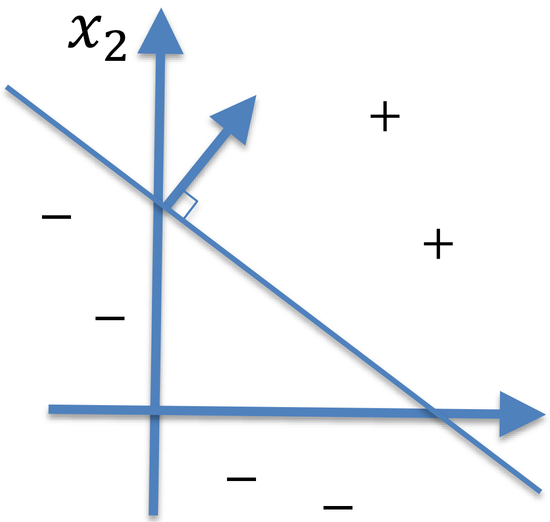
return $\theta^{(k)}$, $\theta_0^{(k)}$

$$\alpha_i^{(k)} = \#\, mistakes\ made\ on\ (x^{(i)}, y^{(i)})\ after\ exactly\ k\ mistakes$$

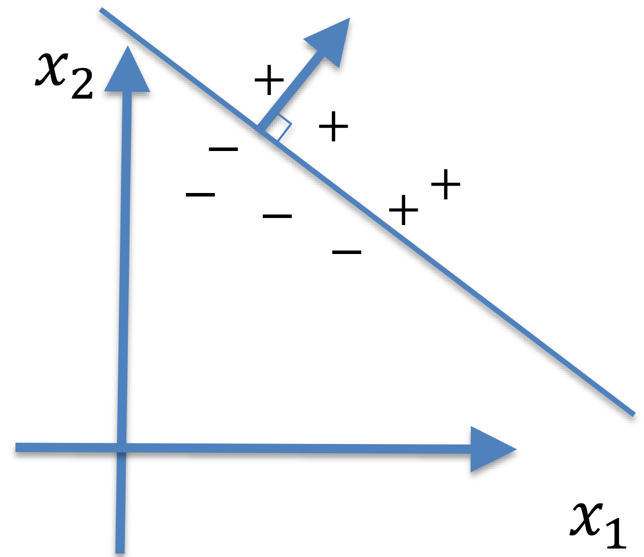
$$\theta^{(k)} = \alpha_1 y^{(1)} x^{(1)} + \cdots + \alpha_n y^{(n)} x^{(n)}$$

$$\theta_0^{(k)} = \alpha_1 y^{(1)} + \cdots + \alpha_n y^{(n)}$$

Convergence



faster convergence

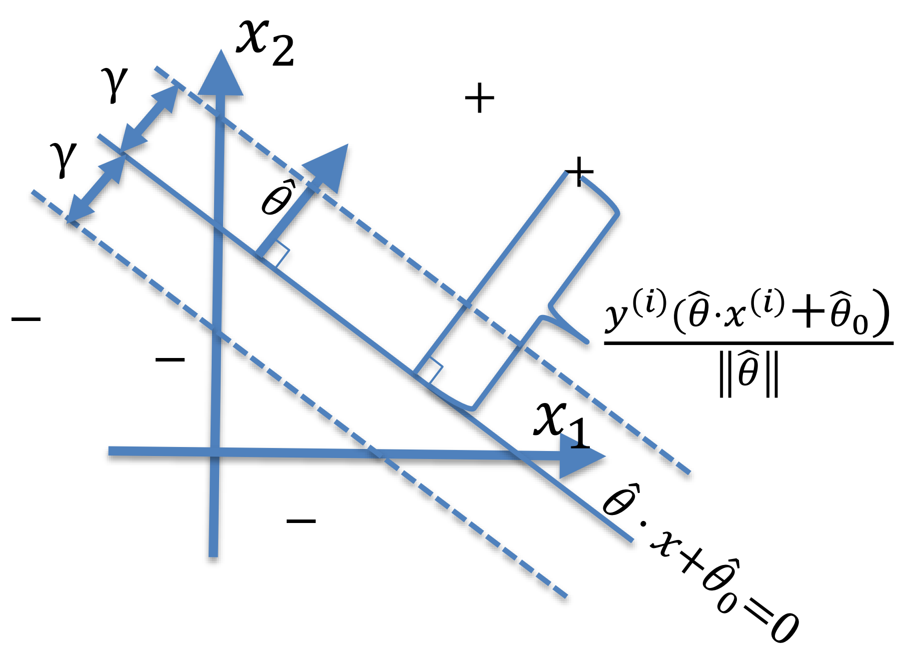


slower convergence

Linear separation with margin

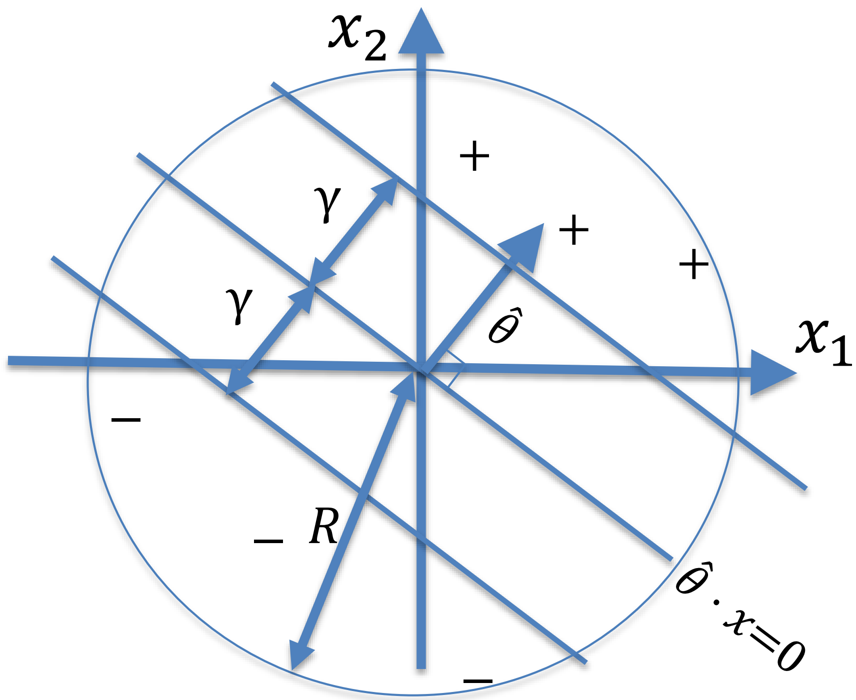
Training examples $S_n = \{(x^{(i)}, y^{(i)}), i = 1 \dots n\}$ are linearly seaparable with margin γ if $\exists \hat{\theta}, \hat{\theta}_0$

s.t. $$\frac{y^{(i)}(\hat{\theta} \cdot x^{(i)} + \hat{\theta}_0)}{\|\hat{\theta}\|} > \gamma > 0 \text{ for all } i = 1 \dots n$$



$$(A)\ \exists\ \hat{\theta}\ s.t.\ \frac{y^{(i)}\hat{\theta}\cdot x^{(i)}}{\|\hat{\theta}\|} > \gamma > 0,\ i = 1 \ldots n$$

$$(B)\ \|x^{(i)}\| \leq R,\ i = 1..n$$



Theorem: If (A) & (B) hold then the perceptron algorithm makes at most $\frac{R^2}{\gamma^2}$ mistakes.

$$\theta^{(k)} - parameter\ after\ k\ mistakes$$

$$1 \geq \cos\left(\angle\left(\theta^{(k)}, \hat{\theta}\right)\right) = \frac{\theta^{(k)} \cdot \hat{\theta}}{\|\theta^{(k)}\|\|\hat{\theta}\|} \geq \frac{k\gamma}{\|\theta^{(k)}\|} \geq \frac{k\gamma}{\sqrt{kR^2}} = \sqrt{\frac{k\gamma^2}{R^2}}$$

k^{th} mistake happened on $\left(x^{(i)}, y^{(i)}\right)$

$$\frac{\theta^{(k)} \cdot \hat{\theta}}{\|\hat{\theta}\|} = \frac{\left(\theta^{(k-1)} + y^{(i)}x^{(i)}\right) \cdot \hat{\theta}}{\|\hat{\theta}\|} = \frac{\theta^{(k-1)} \cdot \hat{\theta}}{\|\hat{\theta}\|} + \underbrace{\frac{y^{(i)}x^{(i)} \cdot \hat{\theta}}{\|\hat{\theta}\|}}_{\geq \gamma} \geq \mathrm{k}\,\gamma$$

$$\|\theta^{(k)}\|^2 = \left\|\left(\theta^{(k-1)} + y^{(i)}x^{(i)}\right)\right\|^2 \leq kR^2$$