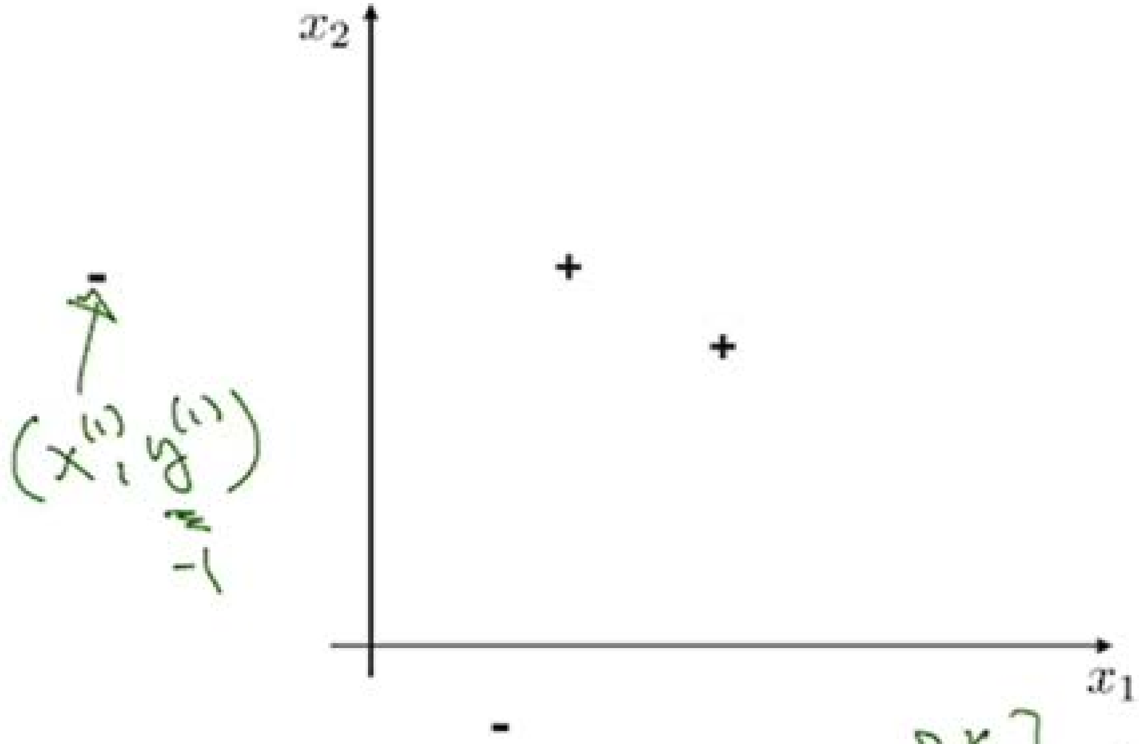
Machine Learning
Lecture 2

Review of basic concepts

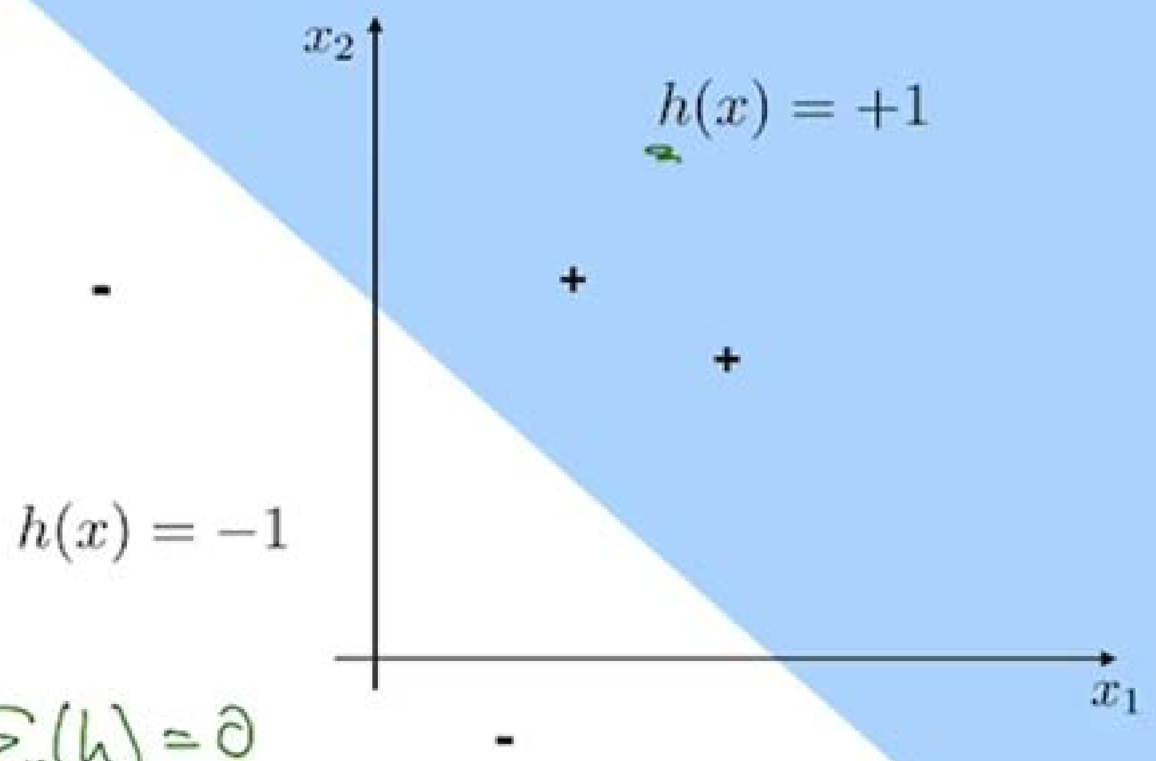
- Feature vectors, labels $x \in \mathbb{R}^d, \ y \in \{-1, 1\}$
- Training set $S_n = \{ (x^{(i)}, y^{(i)}), \ i = 1, \dots, n \}$
- Classifier $h: \mathbb{R}^d \to \{-1, 1\}, \ h(x) = 1, \ \mathcal{X}^+ = \{ x \in \mathbb{R}^d : h(x) = 1 \}$ $\mathcal{X}^-$
- Training error $\mathcal{E}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{[[h(x^{(i)}) \neq y^{(i)}]]}_{= 1 \text{ if error}, \ 0 \text{ o.w.}}$
- Test error $\mathcal{E}(h)$
- Set of classifiers $h \in \mathcal{H}$

Review: training set



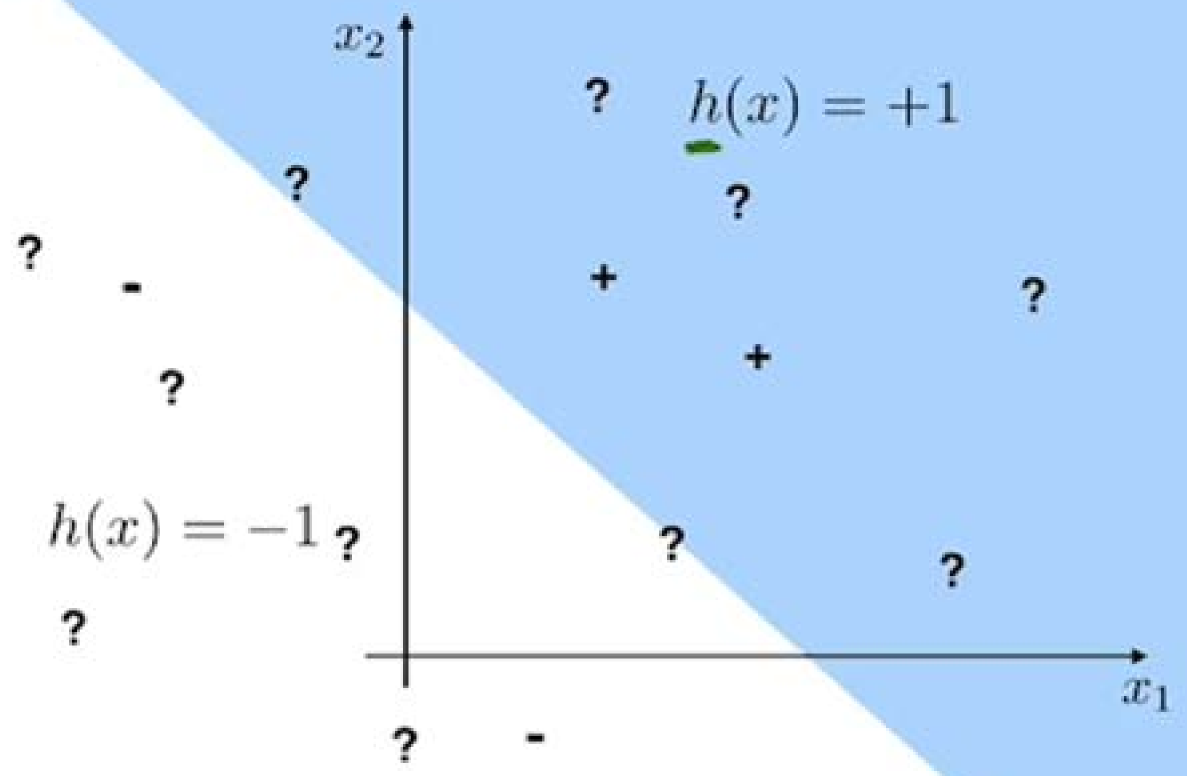
$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^d, \quad d = 2$$

Review: a classifier



$\mathcal{E}_n(h) = 0$

Review: test set



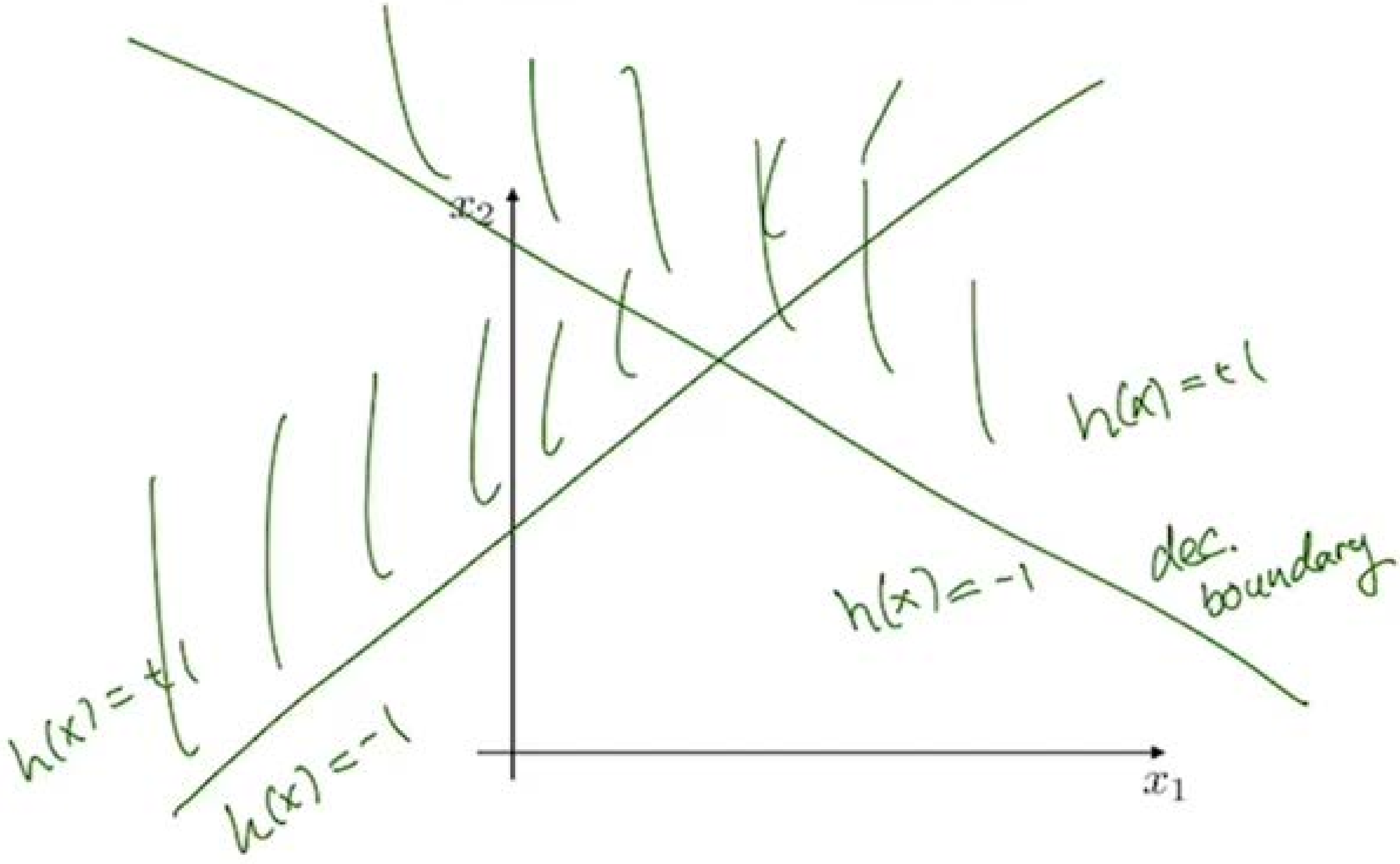
This lecture

- The set of linear classifiers $\mathcal{H}$
- Linear separation
- Perceptron algorithm

$\hat{h} = \mathcal{A}(S_n, \mathcal{H})$

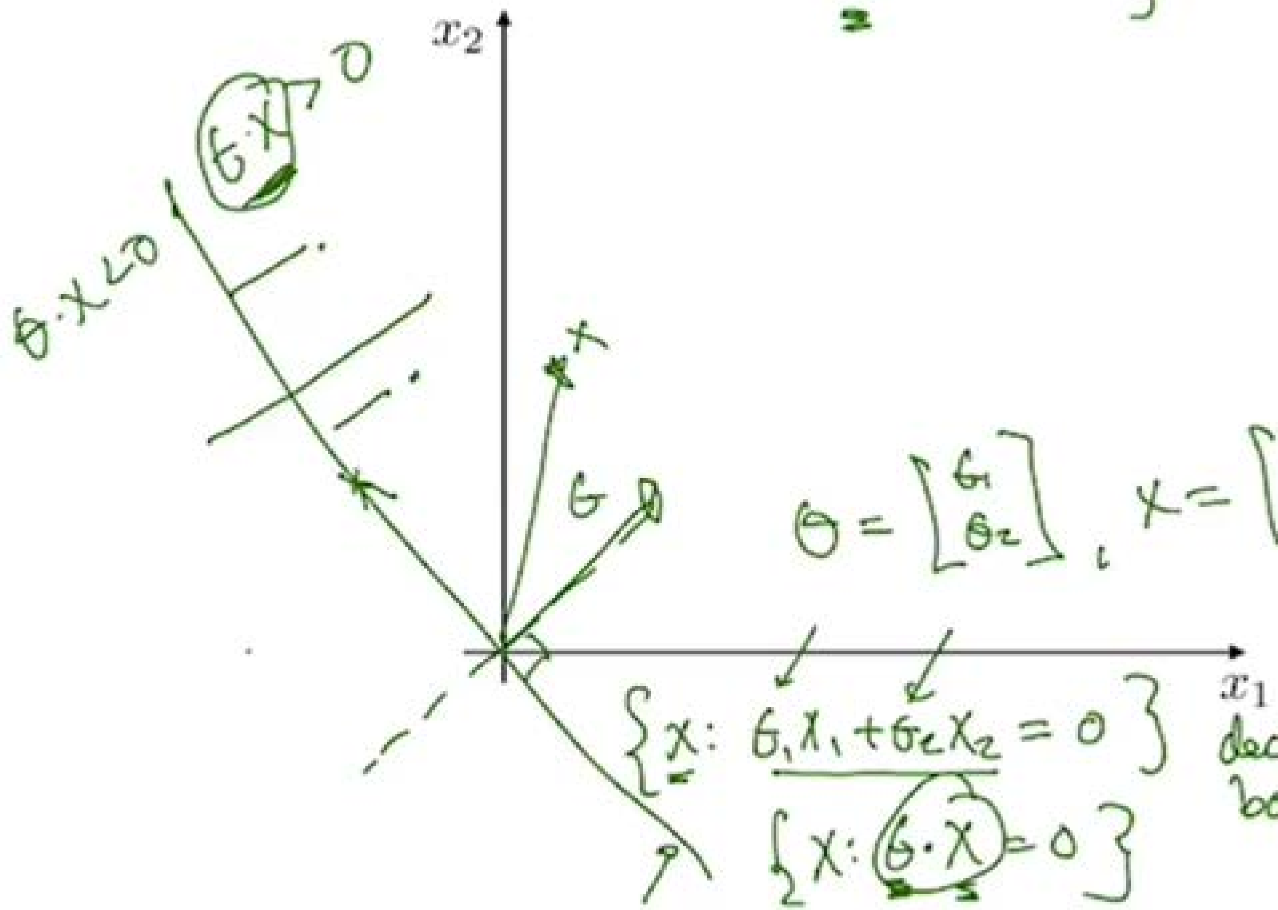
$\hat{h}(x)$

Linear classifiers



Linear classifiers through origin

$$\left\{ h(x;\theta) = \text{sign}(\theta \cdot x), \quad \theta \in \mathbb{R}^d \right\}$$



$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\{x : \theta_1 x_1 + \theta_2 x_2 = 0\}$$ decision boundary

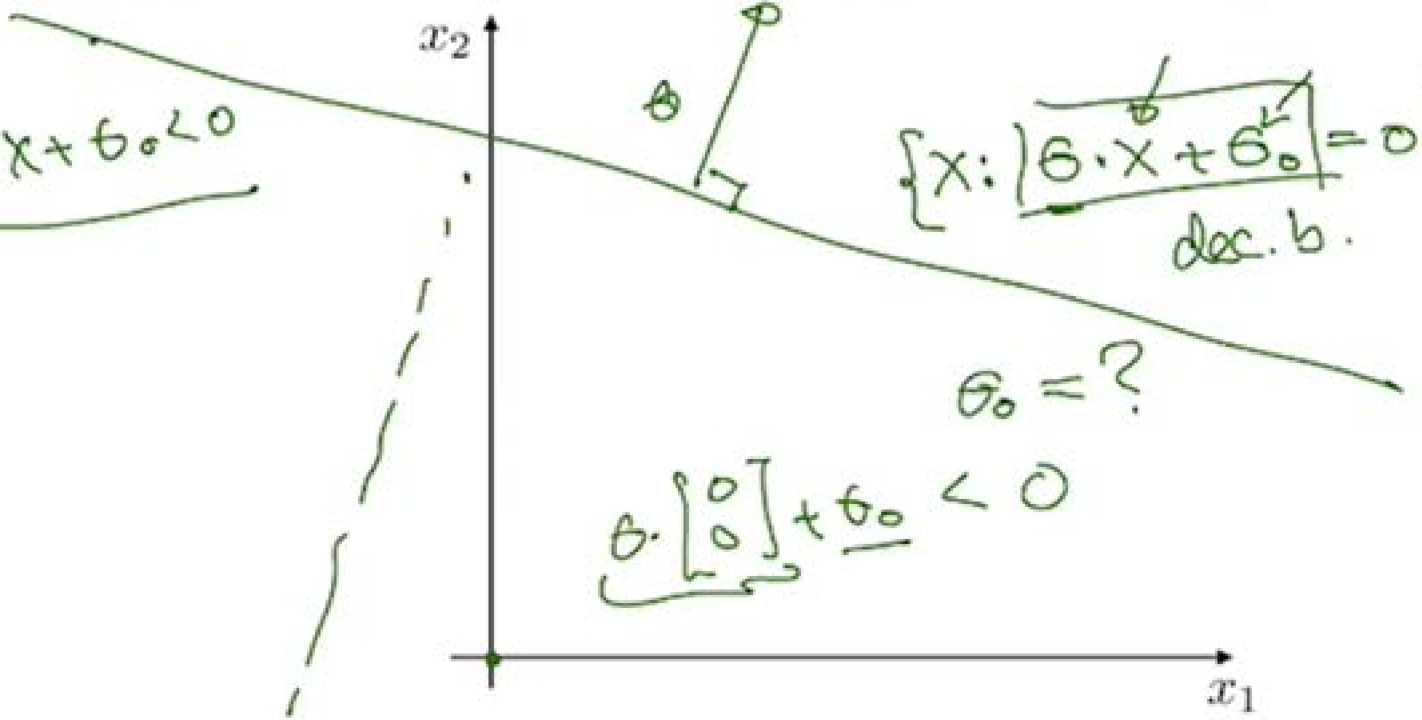
$$\{x : \theta \cdot x = 0\}$$

Linear classifiers

$$\left\{ h(x;\theta,\theta_0) = \text{sign}(\theta\cdot x+\theta_0),\quad \theta\in\mathbb{R}^d,\ \theta_0\in\mathbb{R} \right\}$$

$\theta\cdot x+\theta_0>0$

$\theta\cdot x+\theta_0<0$



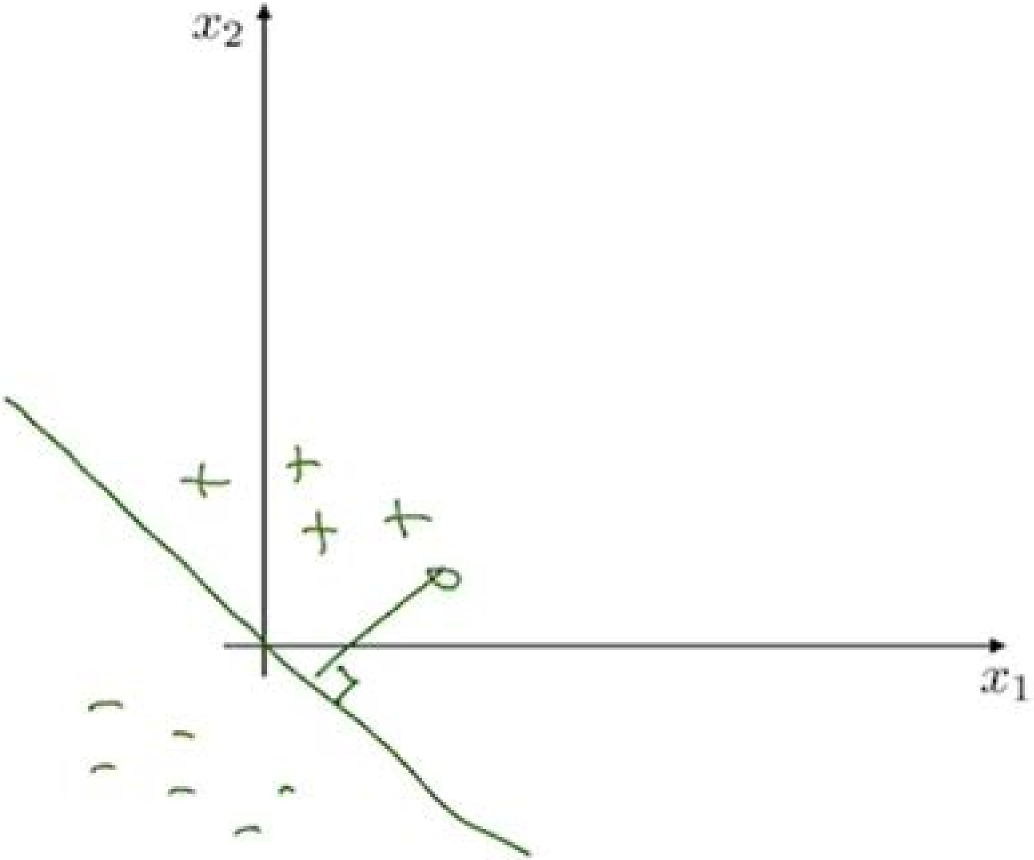
$\{x : \boxed{\theta\cdot x+\theta_0} = 0\}$

dec. b.

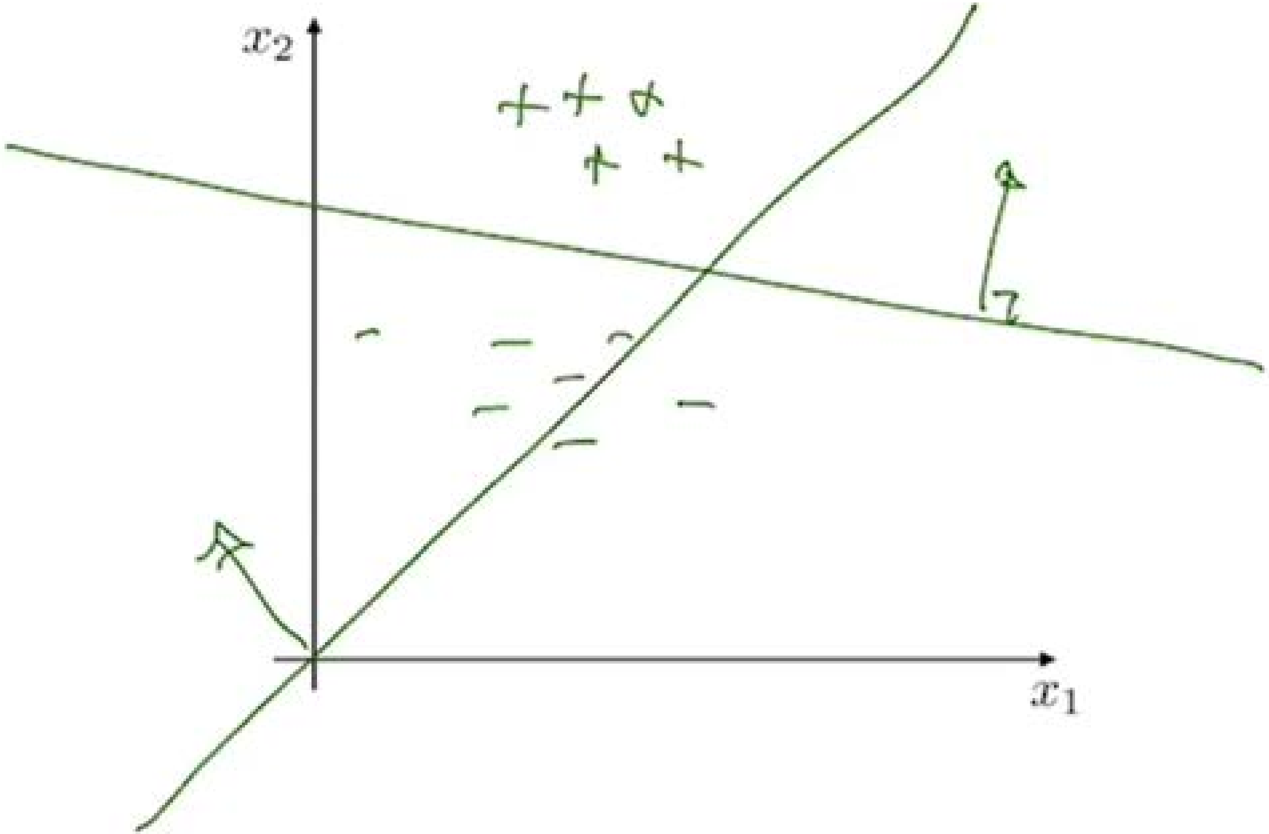
$\theta_0 = ?$

$\theta\cdot\begin{bmatrix}0\\0\end{bmatrix}+\theta_0<0$

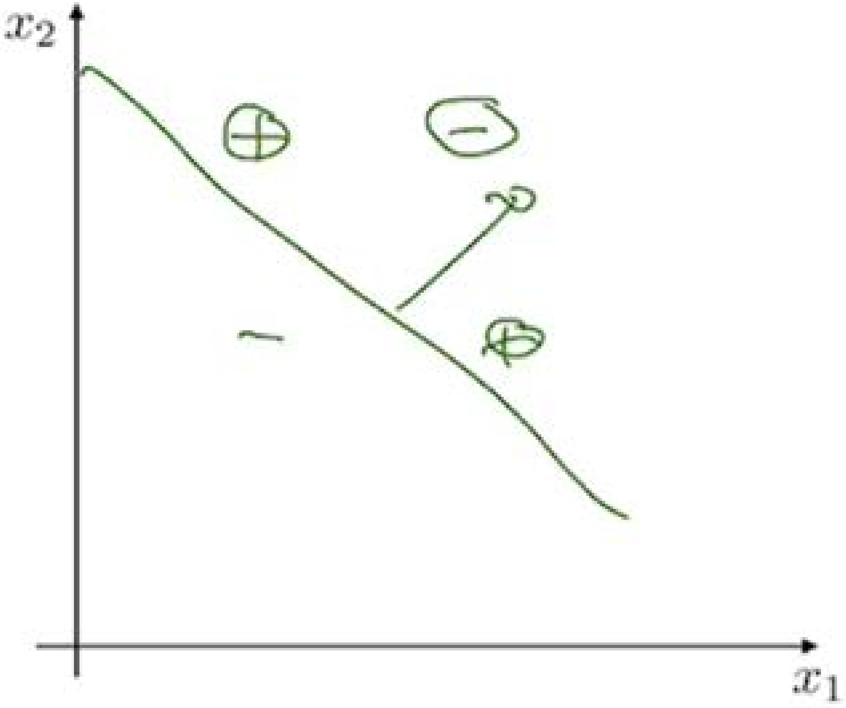
Linear separation: ex



Linear separation: ex



Linear separation: ex



Linear separation

Definition:
Training examples $S_n = \{(x^{(i)}, y^{(i)}\}), i = 1, \ldots, n\}$ are *linearly separable* if there exists a parameter vector $\hat{\theta}$ and offset parameter $\hat{\theta}_0$ such that $y^{(i)}(\hat{\theta} \cdot x^{(i)} + \hat{\theta}_0) > 0$ for all $i = 1, ..., n$.

Learning linear classifiers

- Training error for a linear classifier (through origin)

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

$$\mathcal{E}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} [\![\underline{y^{(i)}\,\theta\cdot x^{(i)}} \leq 0]\!]$$

Learning linear classifiers

- Training error for a linear classifier

$$\mathcal{E}_n(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} [\![y^{(i)} (\theta \cdot x^{(i)} + \theta_0) \leq 0]\!]$$

Learning algorithm: perceptron

$\theta = 0$ (vector)

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

$\theta = \theta + y^{(i)} x^{(i)}$

$y^{(i)}(\cancel{\theta} + y^{(i)} x^{(i)}) \cdot x^{(i)} = \|x^{(i)}\|^2 > 0$

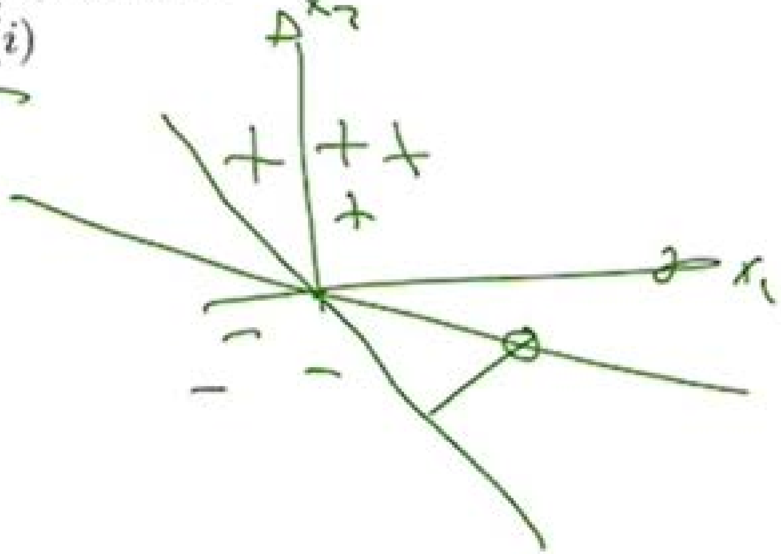
Learning algorithm: perceptron

$\theta = 0$ (vector)

for $i = 1, \ldots, n$ **do**
  if $y^{(i)}(\theta \cdot x^{(i)}) \leq 0$ **then**
    $\theta = \theta + y^{(i)} x^{(i)}$

Learning algorithm: perceptron

procedure PERCEPTRON($\{(x^{(i)}, y^{(i)}), i = 1, \ldots, n\}, T$)
    $\theta = 0$ (vector)
    for $t = 1, \ldots, T$ **do**
        for $i = 1, \ldots, n$ **do**
            if $y^{(i)}(\theta \cdot x^{(i)}) \leq 0$ **then**
                $\theta = \theta + y^{(i)} x^{(i)}$
    return θ



Perceptron algorithm: ex

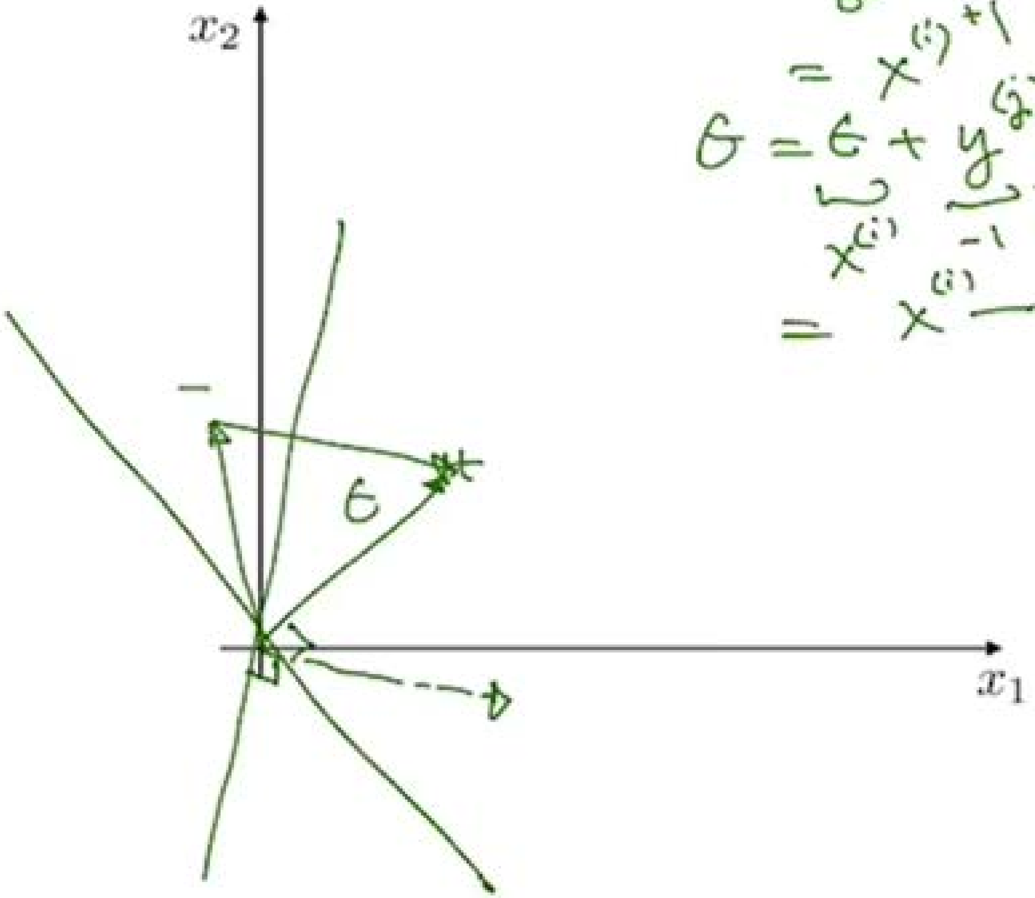
$$\theta = 0$$

$$\theta = \underbrace{\theta}_{0} + \underbrace{y^{(i)}}_{+1} \underbrace{x^{(i)}}$$

$$= x^{(i)}$$

$$\theta = \underbrace{\theta}_{x^{(i)}} + \underbrace{y^{(j)}}_{-1} \underbrace{x^{(j)}}$$

$$= x^{(i)} - x^{(j)}$$



Perceptron (with offset)

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

  $\theta = 0$ (vector), $\theta_0 = 0$ (scalar)

  for $t = 1, \ldots, T$ **do**

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

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

        $\theta = \theta + y^{(i)} x^{(i)}$

        $\theta_0 = \theta_0 + y^{(i)}$

  return θ, θ_0

$$\begin{bmatrix} \theta \\ \theta_0 \end{bmatrix} = \begin{bmatrix} \theta \\ \theta_0 \end{bmatrix} + y^{(i)} \begin{bmatrix} x^{(i)} \\ 1 \end{bmatrix}$$

$$\underbrace{\theta \cdot x + \theta_0} = \underbrace{\begin{bmatrix} \theta \\ \theta_0 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}}$$

Key things to understand

- Parametric families (sets) of classifiers
- The set of linear classifiers
- Linear separation
- Perceptron algorithm