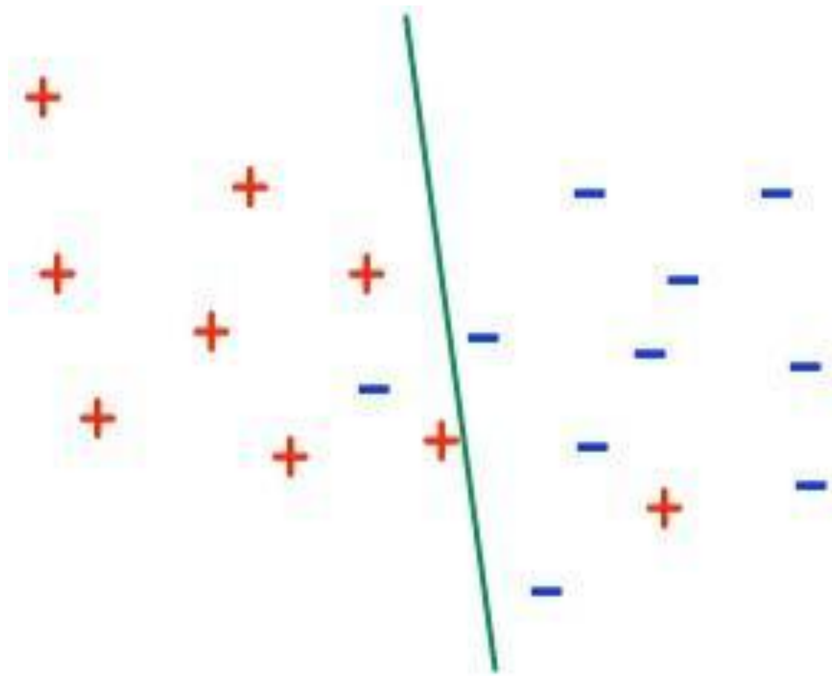
# Perceptron

MGTF 495

# Class Outline

- Generative vs Discriminative Models
- Discriminative Models
    - Logistic Regression
    - SVM
    - Perceptron
- Kernels
- Richer Output Spaces

# The decision boundary



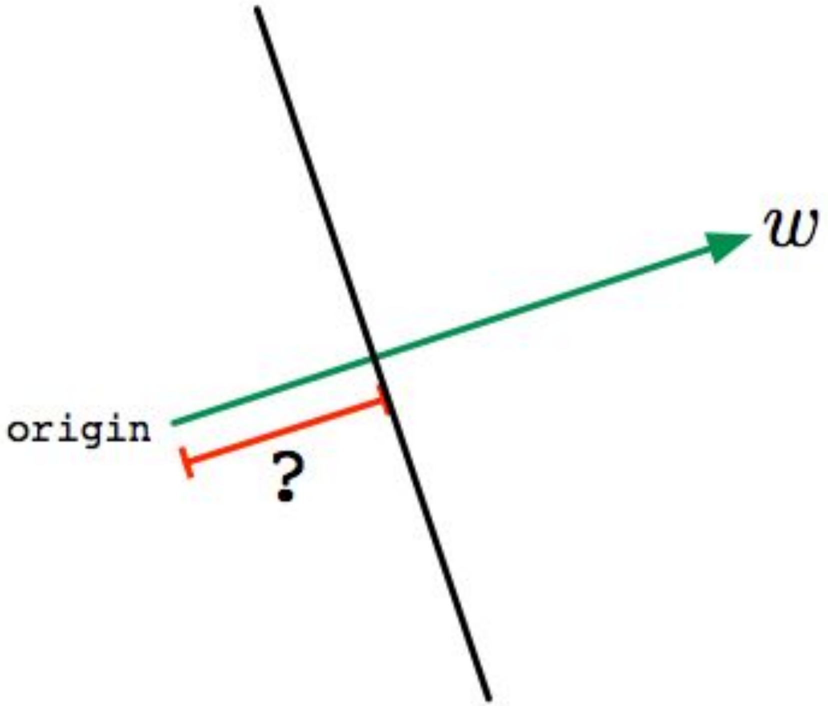

Decision boundary in $R^p$ is a **hyperplane**.

- How is this boundary parametrized?
- How can we learn a hyperplane from training data?

# Hyperplanes

Hyperplane $\{x : w \cdot x = b\}$

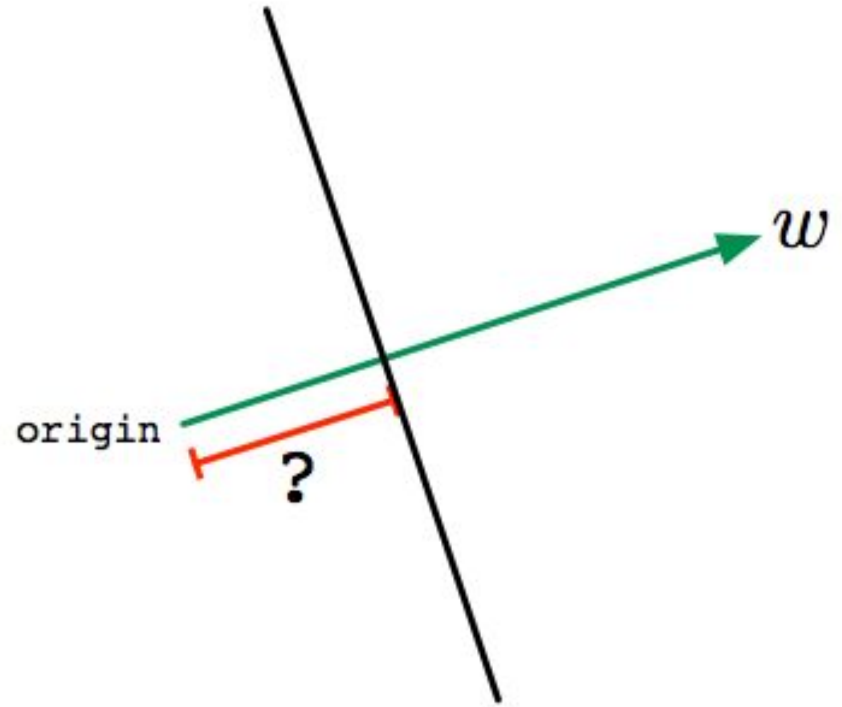
- orientation $w \in R^p$
- offset $b \in R$



# Hyperplanes

Hyperplane $\{x : w \cdot x = b\}$

- orientation $w \in R^p$
- offset $b \in R$




Can always normalize $w$ to unit length:

$$(w, b) \longleftrightarrow \left(\hat{w} = \frac{w}{\|w\|}, \frac{b}{\|w\|}\right)$$

$$w \cdot x = b \longleftrightarrow \hat{w} \cdot x = \frac{b}{\|w\|}$$

Equivalently: all points whose projection onto $\hat{w}$ is $b/\|w\|$.

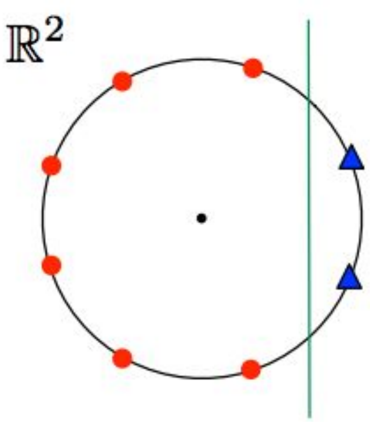
# Homogeneous linear separators

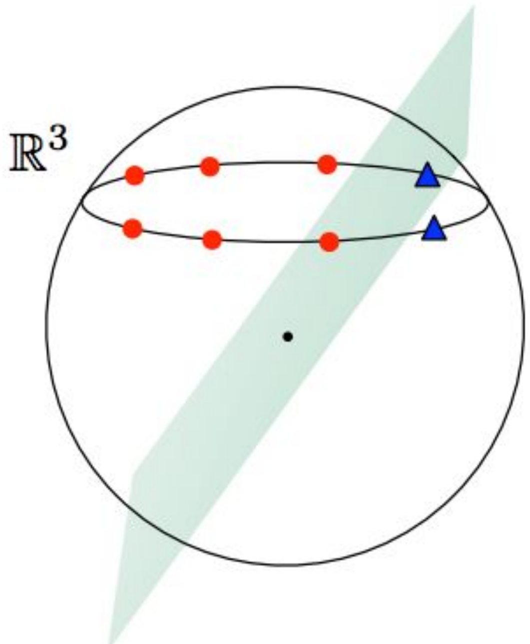
Hyperplanes that pass through the origin have no offset, *b* = 0.

Reduce to this case by adding an extra feature to *x* :

$$\widetilde{x} = (x, 1) \in \mathbb{R}^{p+1}$$

Then $\{x : w \cdot x = b\} \equiv \{x : \widetilde{w} \cdot \widetilde{x} = 0\}$ where $\widetilde{w} = (w, -b)$.



# The learning problem: separable case

*Input:* training data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^p \times \{-1, +1\}$

*Output:* linear classifier $w \in \mathbb{R}^p$ such that

$$y^{(i)}(w \cdot x^{(i)}) > 0 \quad \text{for } i = 1, 2, \ldots, n$$

This is linear programming:

- Each data point is a linear constraint on *w*
- Want to find *w* that satisfies all these constraints

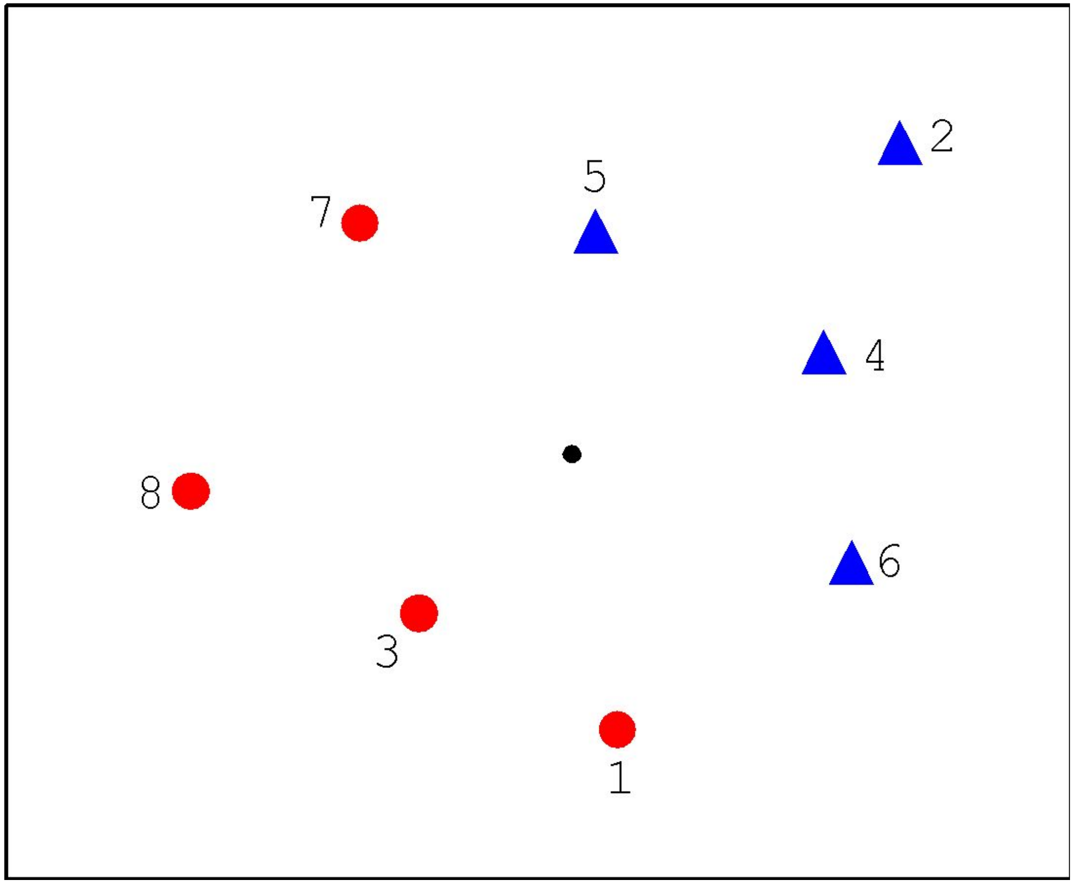
But we won't use generic linear programming methods, such as simplex.

A simple alternative: **Perceptron algorithm** (Rosenblatt, 1958)

- $w = 0$
- while some $(x, y)$ is misclassified:
  - $w = w + yx$

# Perceptron: example

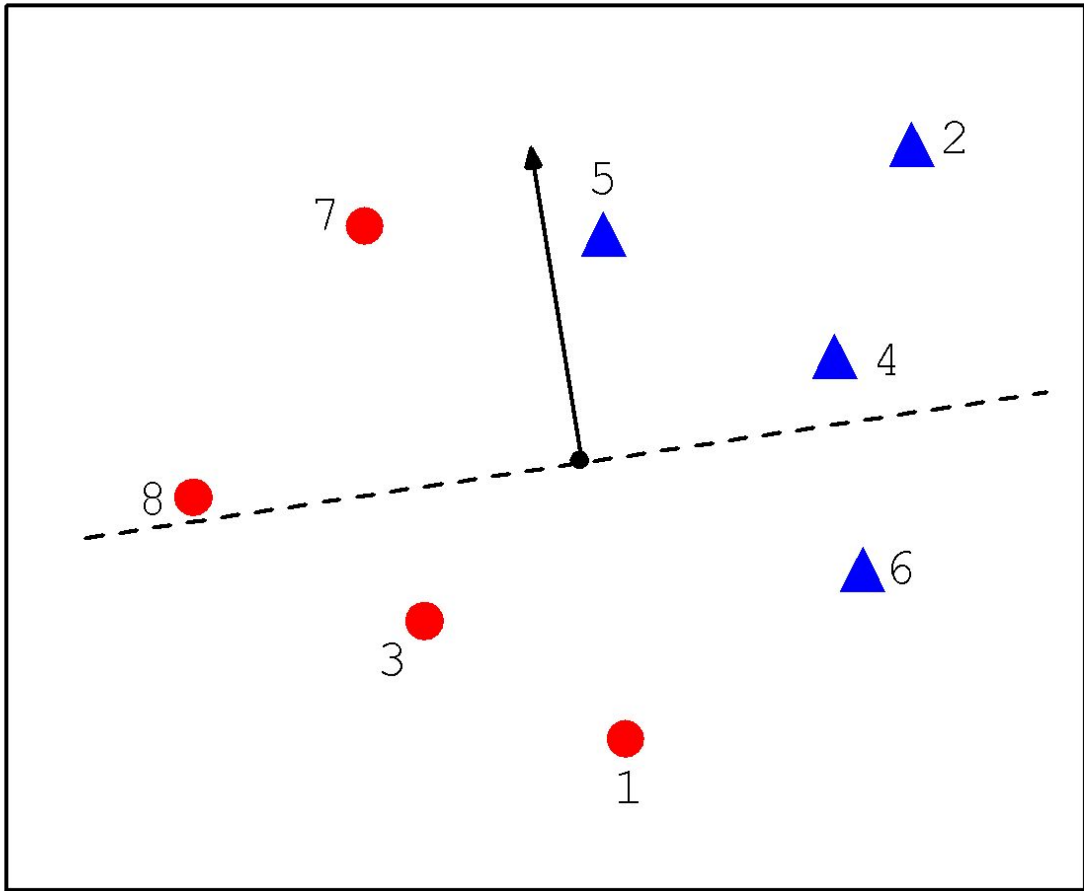
- $w = 0$
- while some $(x , y)$ is misclassified:
    - $w = w + yx$




**Separator:** $w = 0$

# Perceptron: example

- $w = 0$
- while some $(x, y)$ is misclassified:
  - $w = w + yx$

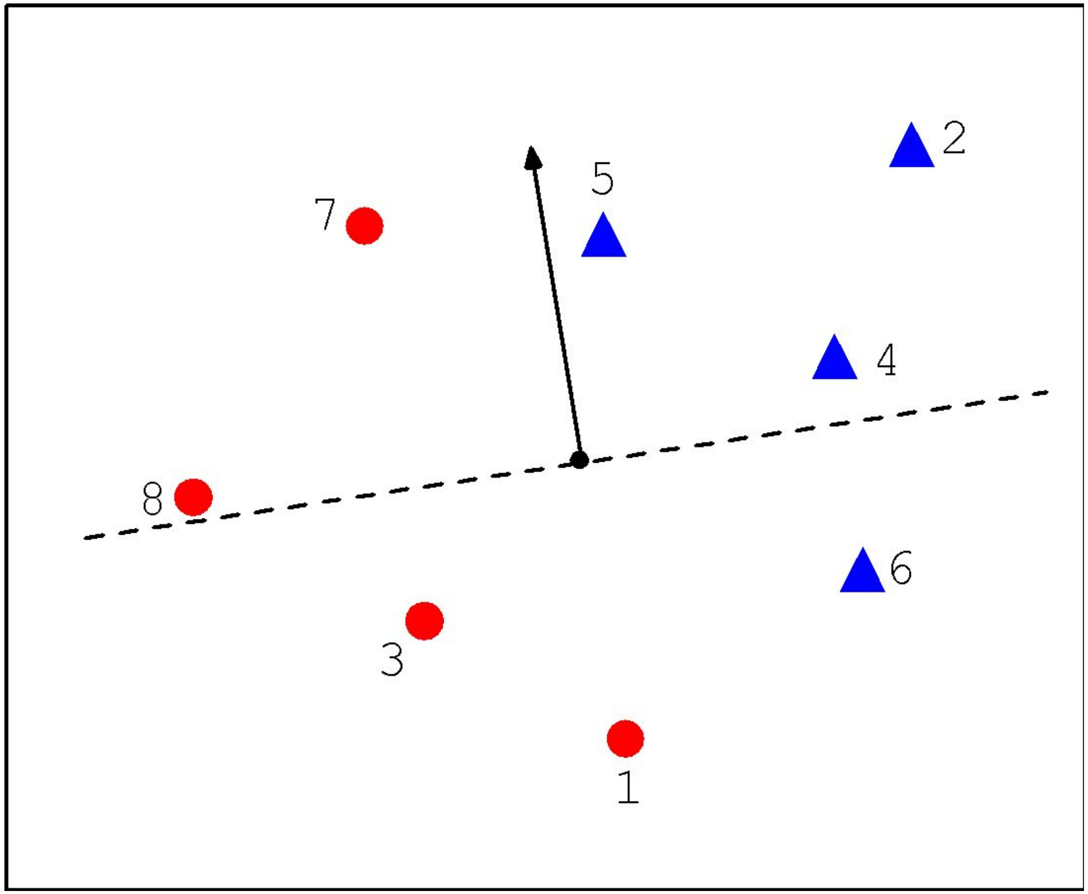



**Separator:** $w = -x^{(1)}$

# Perceptron: example

- $w = 0$
- while some $(x\ ,\ y\ )$ is misclassified:
  - $w = w + yx$

Points 2-5 are correct

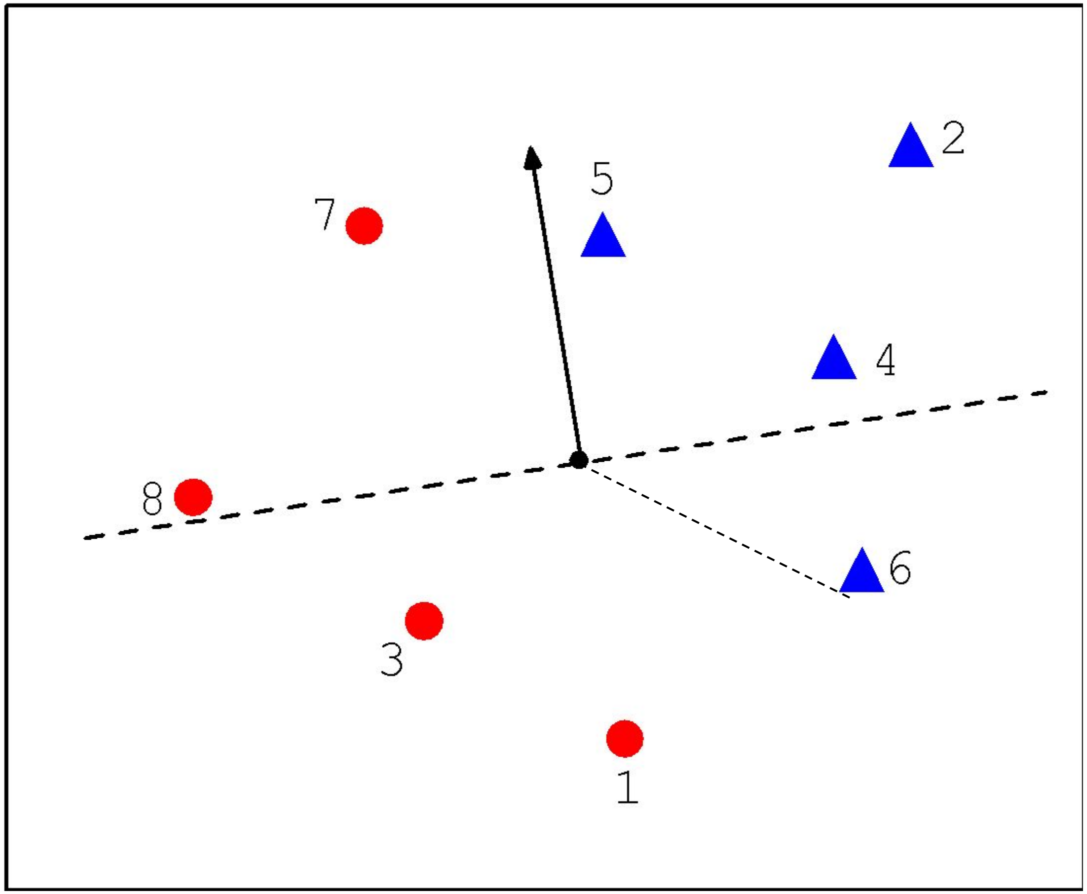



**Separator:** $w = -x^{(1)}$

# Perceptron: example

- $w = 0$
- while some $(x\ ,\ y\ )$ is misclassified:
  - $w = w + yx$

Six is misclassified
-> Add vector in direction of 6

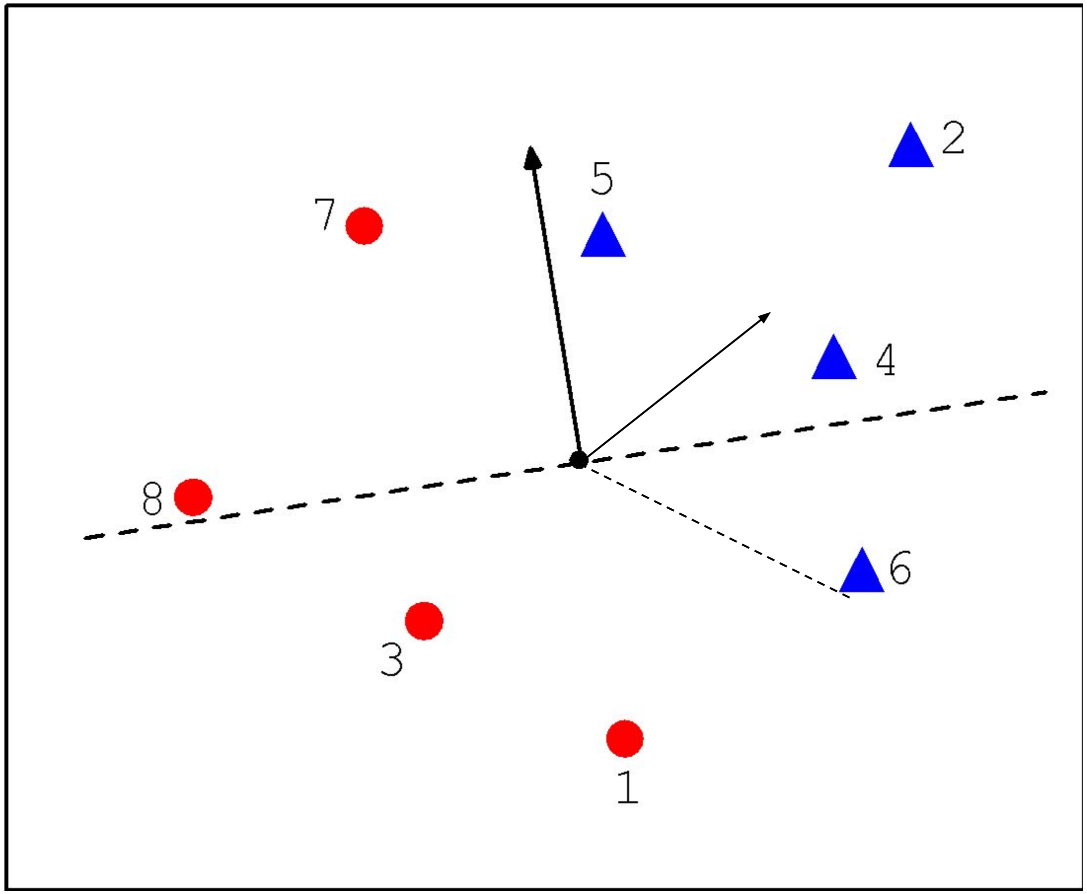



**Separator:** $w = -x^{(1)}$

# Perceptron: example

- $w = 0$
- while some $(x , y)$ is misclassified:
  - $w = w + yx$

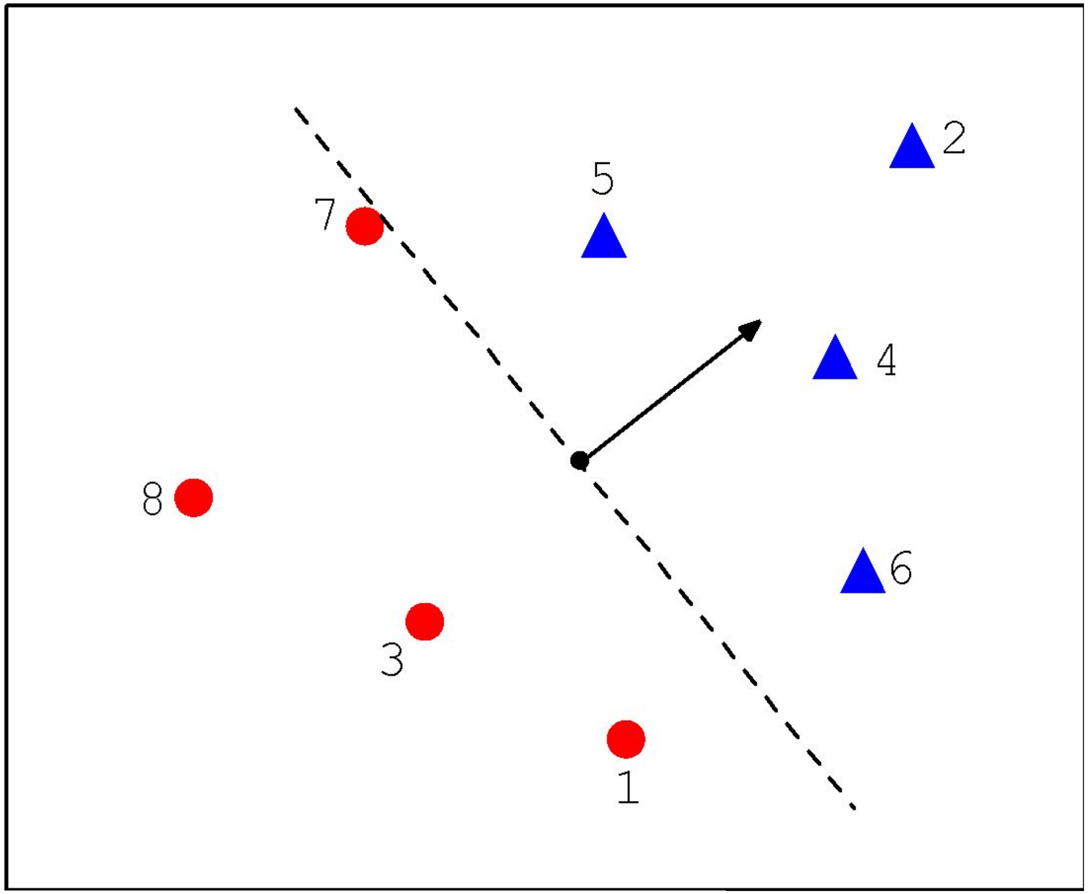
Six is misclassified
-> Add vector in direction of 6




**Separator:** $w = -x^{(1)}$

# Perceptron: example

- $w = 0$
- while some $(x, y)$ is misclassified:
    - $w = w + yx$




**Separator:** $w = 0 w = -x^{(1)} w = -x^{(1)} + x^{(6)}$

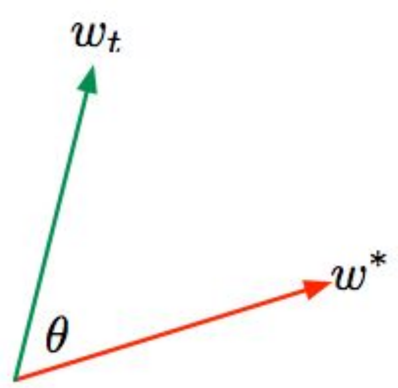
# Perceptron: convergence

**Theorem:** Let $R = \max \|x^{(i)}\|$. Suppose there is a unit vector $w^*$ and some (margin) $\gamma > 0$ such that

$$y^{(i)}(w^* \cdot x^{(i)}) \geq \gamma \quad \text{for all } i.$$

Then the Perceptron algorithm converges after at most $R^2/\gamma^2$ updates.

**Proof idea.** Let $w_t$ be the classifier after $t$ updates.




**Track angle between $w_t$ and $w^*$:**

$$\cos(\angle(w_t, w^*)) = \frac{w_t \cdot w^*}{\|w\|}.$$

On each mistake, when $w_t$ is updated to $w_{t+1}$,

- $w_t \cdot w^*$ grows significantly.
- $\|w_t\|$ does not grow much.

# Perceptron convergence, cont'd

Perceptron update: if $y(w_t \cdot x) < 0$ (misclassified) then $w_{t+1} = w_t + yx$.
Target vector $w^*$ has unit length, and margin condition $y(w^* \cdot x) \geq \gamma$.

1. Initial vector $w_0 = 0$.
2. When updating $w_t$ to $w_{t+1}$:

$$w_{t+1} \cdot w^* = (w_t + yx) \cdot w^* = w_t \cdot w^* + y(w^* \cdot x) \geq w_t \cdot w^* + \gamma$$

$$\|w_{t+1}\|^2 = \|w_t + yx\|^2 = \|w_t\|^2 + \|x\|^2 + 2y(w_t \cdot x) \leq \|w_t\|^2 + R^2$$

3. After $T$ updates, we have

$$w_T \cdot w^* \geq T\gamma$$

$$\|w_T\|^2 \leq TR^2$$

4. The angle between $w_T$ and $w^*$ is given by

$$\cos(\angle(w_T, w^*)) = \frac{w_T \cdot w^*}{\|w\|} \geq \frac{T\gamma}{R\sqrt{T}}.$$

**This is at most 1, so $T \leq R^2/\gamma^2$.**