ONLINE MASTERS IN **DATA SCIENCE**

DSC 255 - MACHINE LEARNING FUNDAMENTALS

MAXIMIZING THE MARGIN OF A LINEAR CLASSIFIER

SANJOY DASGUPTA, PROFESSOR

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Recall: The Perceptron Algorithm

The Perceptron Algorithm

- Initialize $w = 0$ and $b = 0$
- Keep cycling through the training data (x, y):
  - If $y(w \cdot x + b) \leq 0$ (i.e., point misclassified):
    - $w = w + yx$
    - $b = b + y$

Perceptron: Convergence

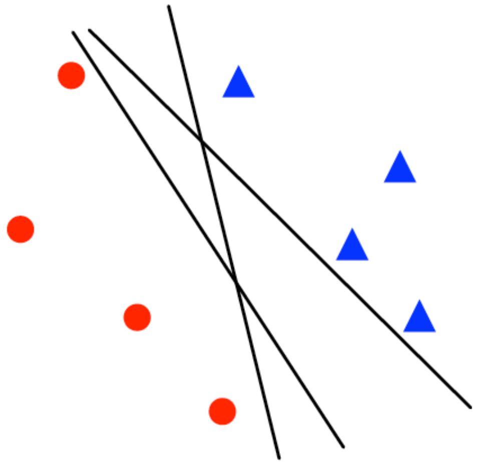
If the training data is linearly separable:

- The Perceptron algorithm will find a linear classifier with zero training error
- It will converge within a finite number of steps

Perceptron: Convergence

If the training data is linearly separable:

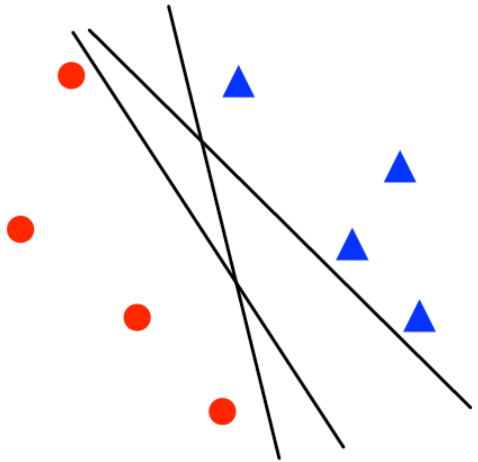
- The Perceptron algorithm will find a linear classifier with zero training error
- It will converge within a finite number of steps



Perceptron: Convergence

If the training data is linearly separable:

- The Perceptron algorithm will find a linear classifier with zero training error
- It will converge within a finite number of steps



Is there a better, more systematic choice of separator?

The Learning Problem

Given: training data $\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right) \in \mathbb{R}^d \times \{-1, +1\}$

Find: $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $y^{(i)}(w \cdot x^{(i)} + b) > 0$ for all i.

The Learning Problem

Given: training data $\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right) \in \mathbb{R}^d \times \{-1, +1\}$

Find: $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that $y^{(i)}(w \cdot x^{(i)} + b) > 0$ for all i.

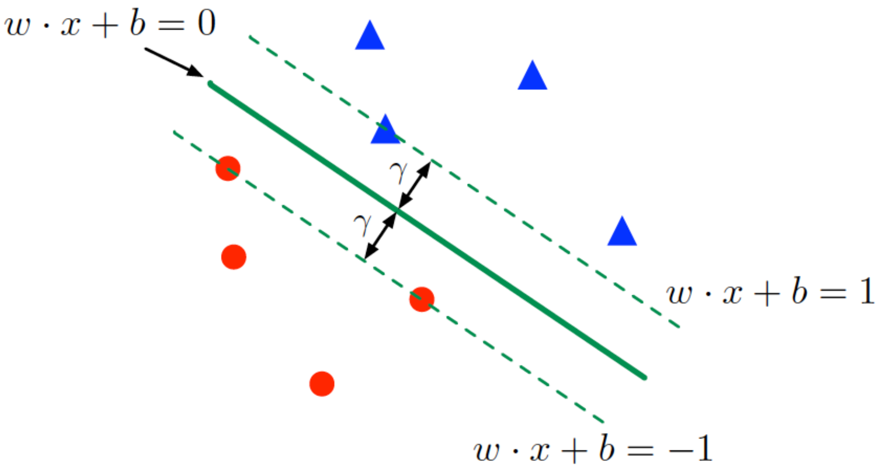
By scaling w, b, can equivalently ask for

$$y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \text{ for all } i$$

Maximizing the Margin

Given: training data $\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right) \in \mathbb{R}^d \times \{-1, +1\}$
Find: $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that

$$y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \text{ for all } i$$

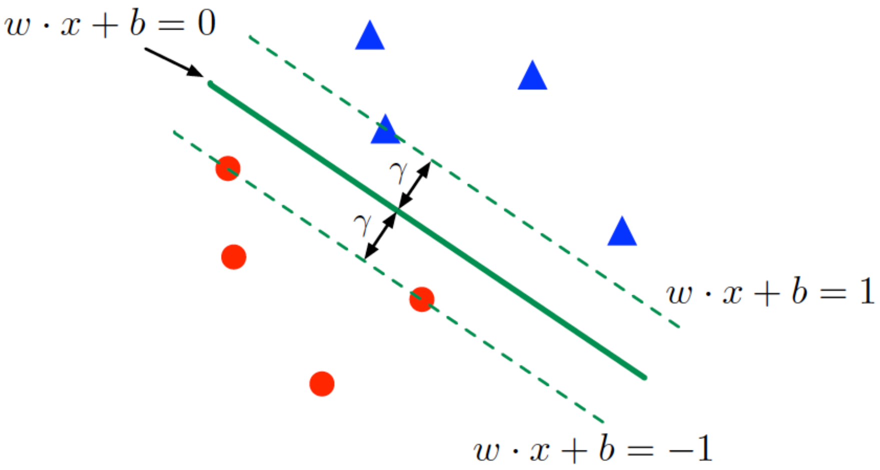


Maximize the **margin** γ.

Maximizing the Margin

Given: training data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$
Find: $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ such that

$$y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \text{ for all } i$$



Maximize the **margin** γ.   Can show $\gamma = 1/\|w\|$.

Maximum-Margin Linear Classifier

- *Given:* $\left(x^{(1)}, y^{(1)}\right), \dots, \left(x^{(n)}, y^{(n)}\right) \in \mathbb{R}^d \times \{-1, +1\}$

$$\boxed{\begin{array}{c} \min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \|w\|^2 \\ \\ y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i = 1, 2, \dots, n \end{array}}$$

Maximum-Margin Linear Classifier

- *Given:* $\left(x^{(1)}, y^{(1)}\right), \dots, \left(x^{(n)}, y^{(n)}\right) \in \mathbb{R}^d \times \{-1, +1\}$

$$\boxed{\begin{array}{c} \min\limits_{w \in \mathbb{R}^d, b \in \mathbb{R}} \|w\|^2 \\ y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i = 1, 2, \dots, n \end{array}}$$

- **Convex optimization problem**: can find the optimal solution efficiently.

Maximum-Margin Linear Classifier

- *Given*: $\left(x^{(1)}, y^{(1)}\right), \dots, \left(x^{(n)}, y^{(n)}\right) \in \mathbb{R}^d \times \{-1, +1\}$

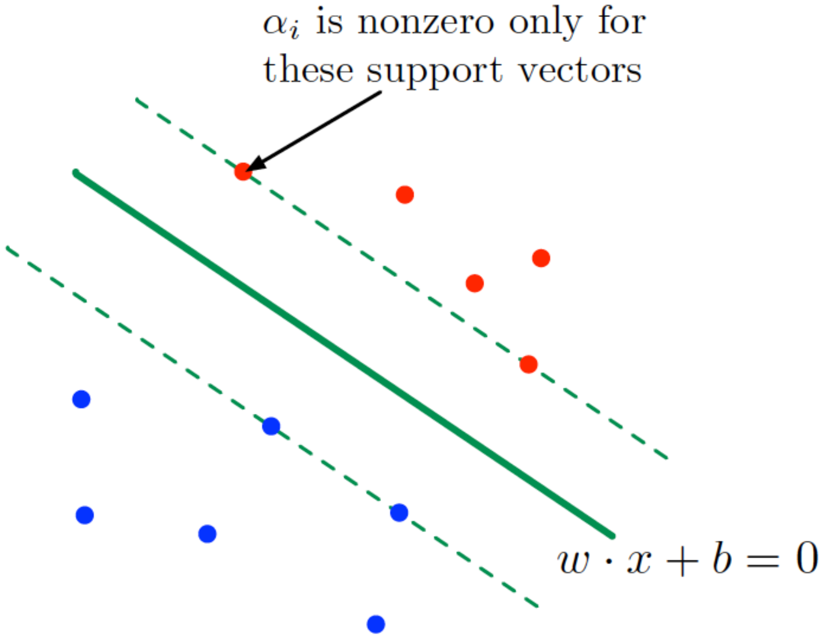
$$\boxed{\begin{gathered}\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \|w\|^2 \\ y^{(i)}(w \cdot x^{(i)} + b) \geq 1 \quad \text{for all } i = 1, 2, \dots, n\end{gathered}}$$

- **Convex optimization problem**: can find the optimal solution efficiently.

- This linear classifier is sometimes called the **(hard-margin) support vector machine.**

Support Vectors

- The solution $w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}$ is a function of just the **support vectors**: training points exactly on the margin, i.e., $y^{(i)}\left(w \cdot x^{(i)} + b\right) = 1$



Small Example: Iris Data Set

Fisher's **iris** data



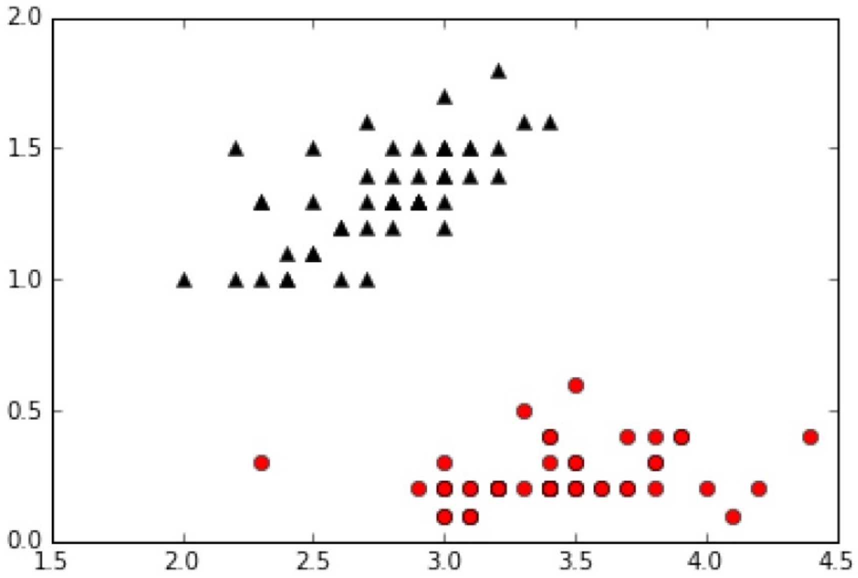
150 data points from three classes:

- iris setosa
- iris versicolor
- iris virginica

Four measurements: petal width/length, sepal width/length

Small Example: Iris Data Set

Two features: sepal width, petal width.
Two classes: setosa (red circles), versicolor (black triangles)



Small Example: Iris Data Set

Two features: sepal width, petal width.
Two classes: setosa (red circles), versicolor (black triangles)