

CS 66: Machine Learning

Prof. Sara Mathieson

Spring 2019



Outline for March 22

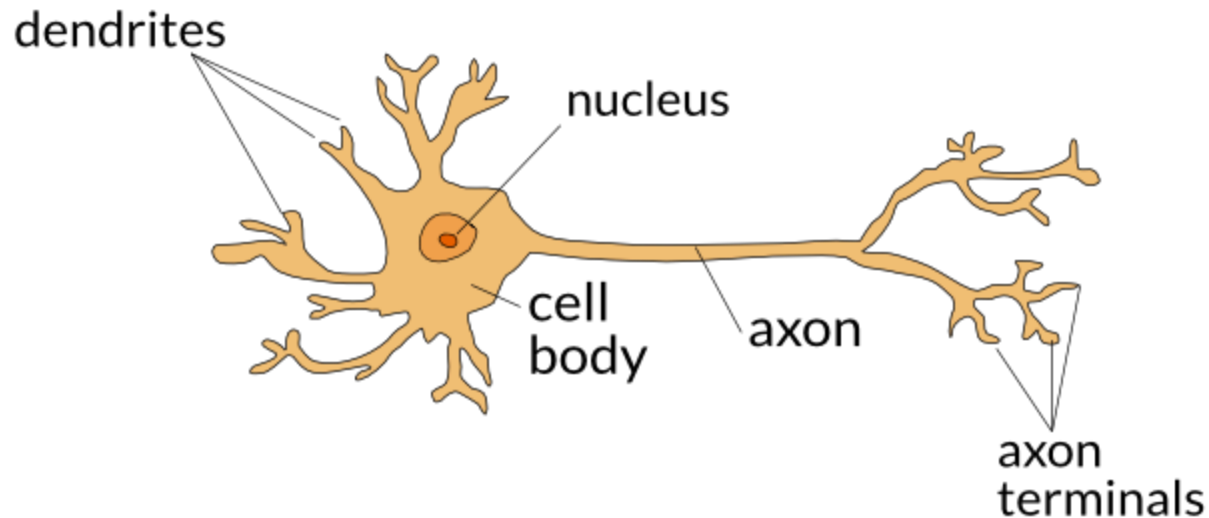
- Perceptron history and interpretation as a neural network
- Idea of a maximum margin classifier
- Support Vector Machines introduction
- Functional vs. Geometric margins
- SVM as an optimization problem
 - Office hours TODAY: 12:30-2:30pm (shift half-hour earlier)
 - Lab 5 due Tuesday night

Outline for March 22

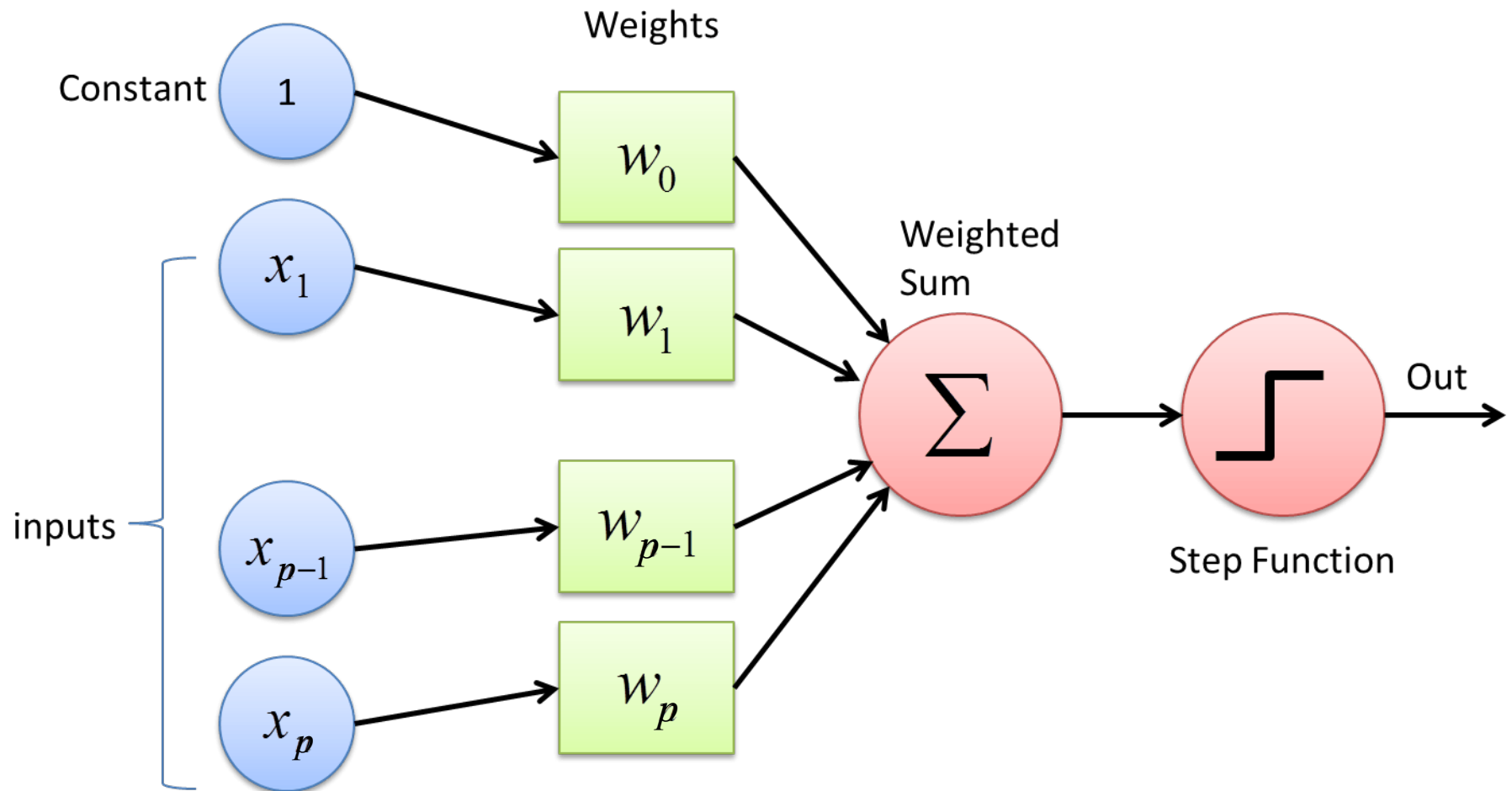
- Perceptron history and interpretation as a neural network
- Idea of a maximum margin classifier
- Support Vector Machines introduction
- Functional vs. Geometric margins
- SVM as an optimization problem

Perceptron as a neural network

Biological model of a neuron



Perceptron as a neural network



History of the Perceptron

- Invented in 1957 by Frank Rosenblatt
- Initially thought to be the “solution to AI”

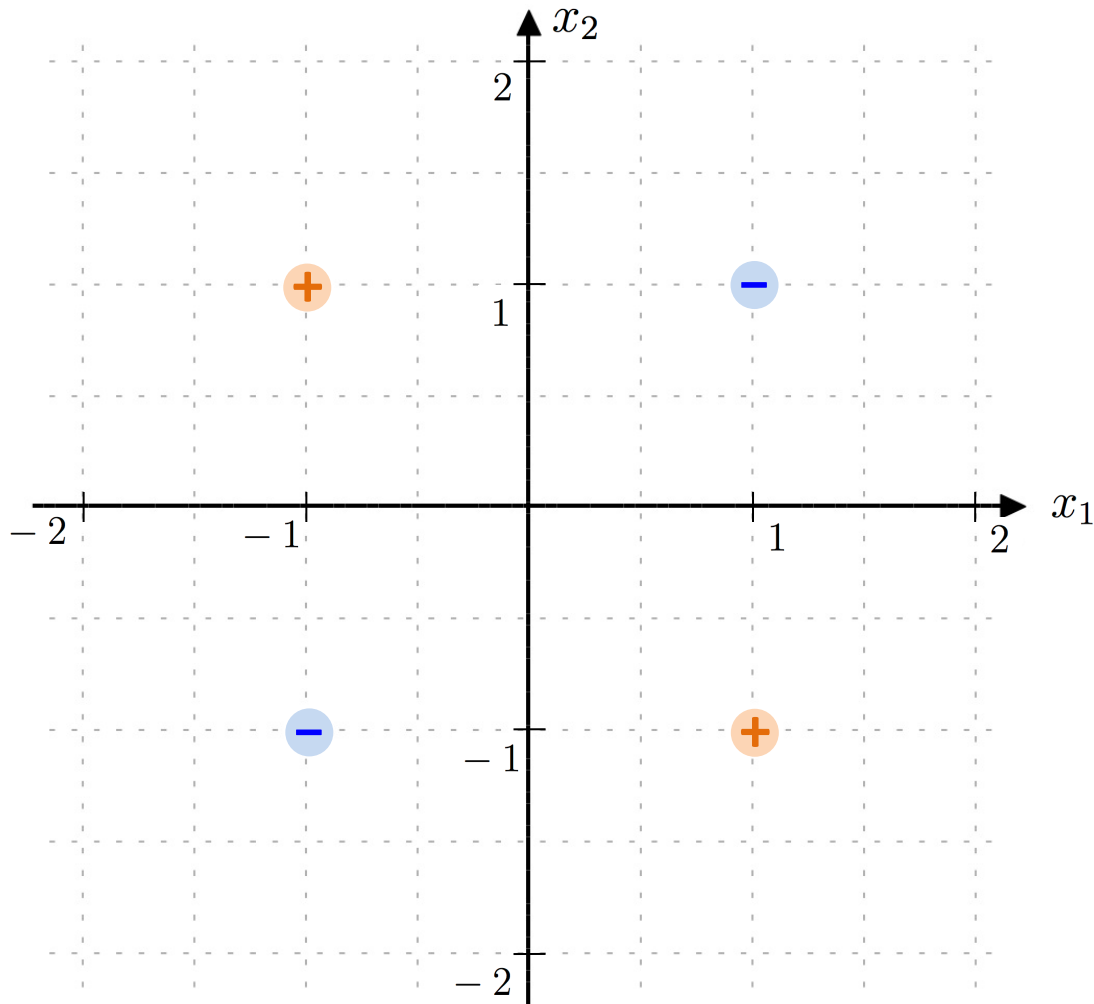
NYT said the perceptron was “*the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence*”

- Famous book “Perceptrons” by Marvin Minsky and Seymour Papert (1969)
- Confusion about the text contributed to first “AI winter”

Perceptron cannot learn XOR

($x_1 = 1$ or $x_2 = 1$, but not both)

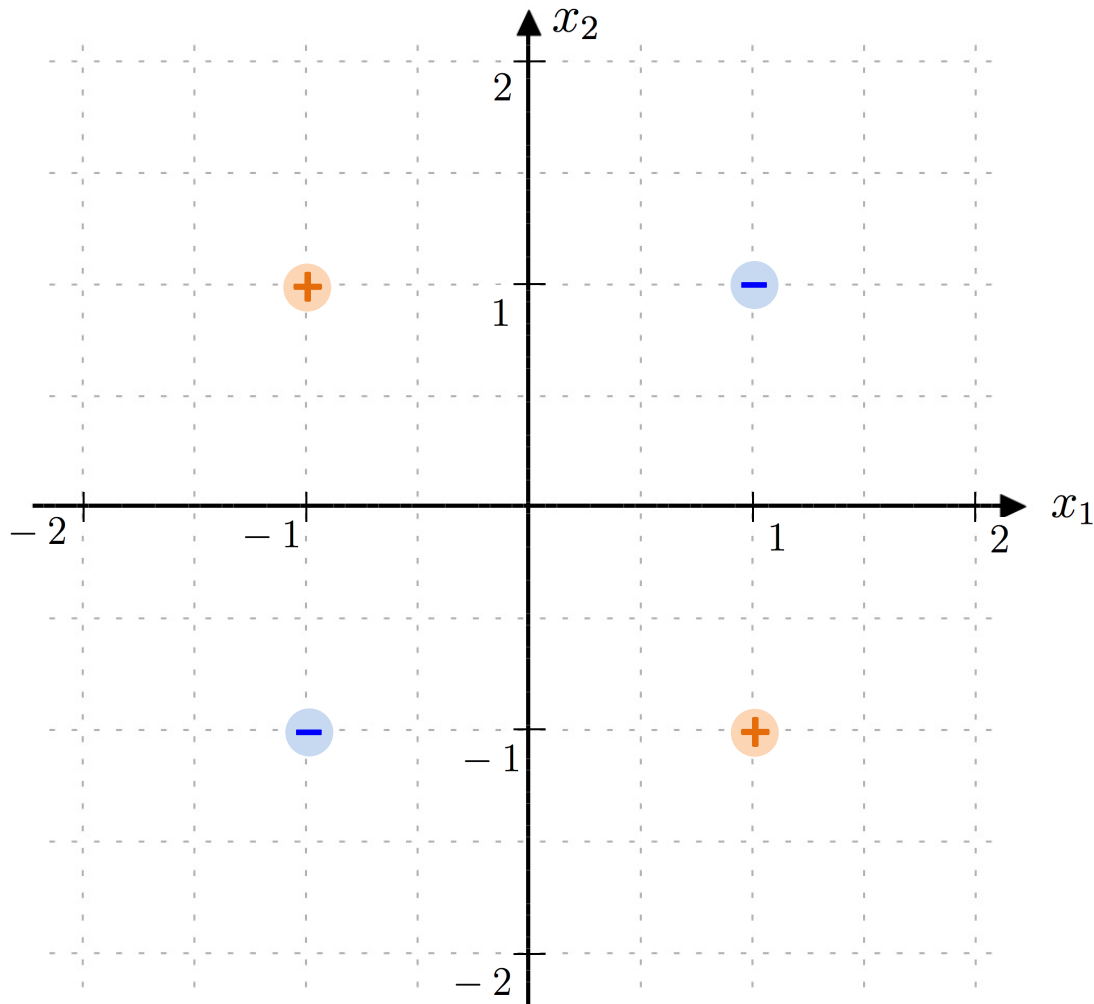
Why?



Perceptron cannot learn XOR

($x_1 = 1$ or $x_2 = 1$, but not both)

Why?
Not linearly
separable!



Handout 10 example

Final solution (so you can check your work):

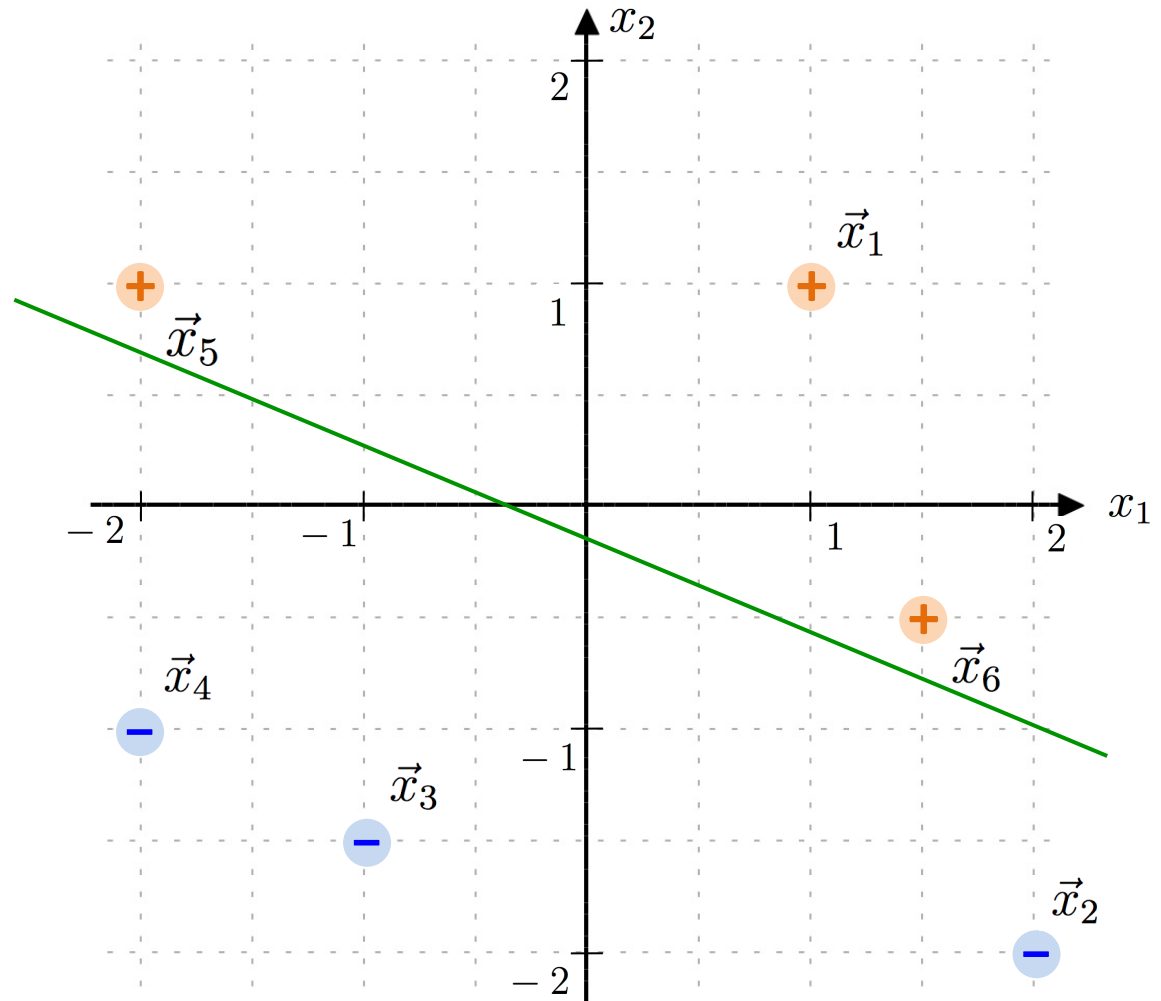
$$\vec{w}^* = \begin{bmatrix} 0.2 \\ 0.5 \\ 1 \end{bmatrix}$$

Final hyperplane:

$$0.2 + 0.5x_1 + x_2 = 0$$

\Rightarrow

$$x_2 = -0.2 - 0.5x_1$$

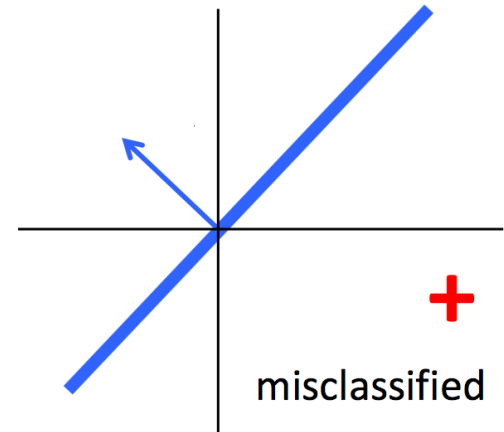


Informal quiz (discuss with a partner)

- 1) What is the relationship between the weight vector \mathbf{w} and the hyperplane?
- 2) Why is the perceptron cost function intuitive?

$$J(\vec{w}) = \sum_{i=1}^n \max \left(0, -y_i (\vec{w}^T \vec{x}_i) \right)$$

- 3) In the example to the right, how will the slope of the hyperplane change?



- 4) What are the weaknesses of the perceptron?
Create a binary classifier “wishlist”.

Informal quiz (discuss with a partner)

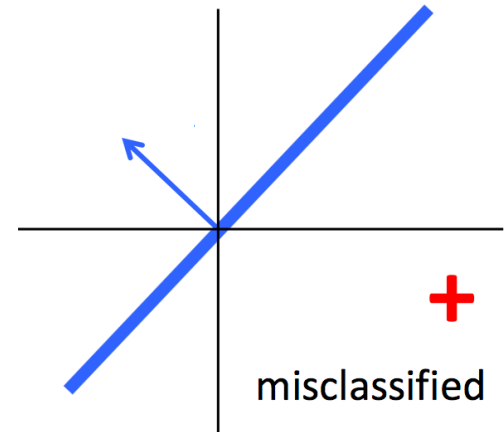
- 1) What is the relationship between the weight vector \mathbf{w} and the hyperplane?

They are perpendicular

- 2) Why is the perceptron cost function intuitive?

$$J(\vec{w}) = \sum_{i=1}^n \max \left(0, -y_i (\vec{w}^T \vec{x}_i) \right)$$

- 3) In the example to the right, how will the slope of the hyperplane change?



- 4) What are the weaknesses of the perceptron?
Create a binary classifier “wishlist”.

Informal quiz (discuss with a partner)

- 1) What is the relationship between the weight vector \mathbf{w} and the hyperplane?

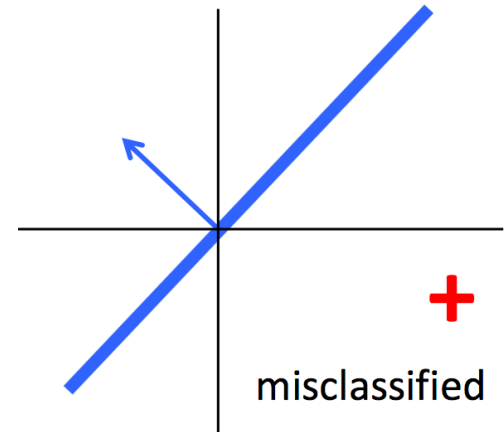
They are perpendicular

- 2) Why is the perceptron cost function intuitive?

$$J(\vec{w}) = \sum_{i=1}^n \max \left(0, -y_i (\vec{w}^T \vec{x}_i) \right)$$

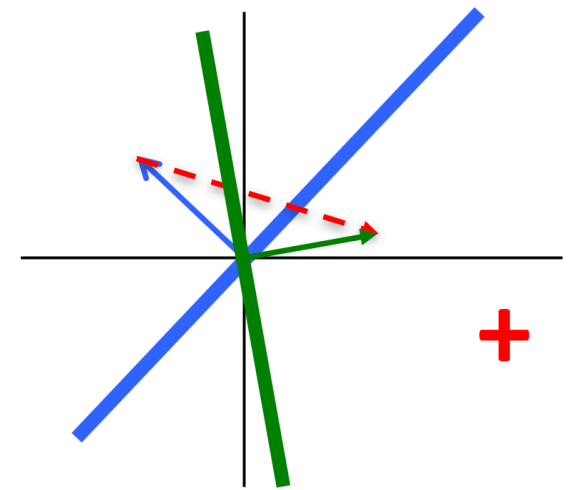
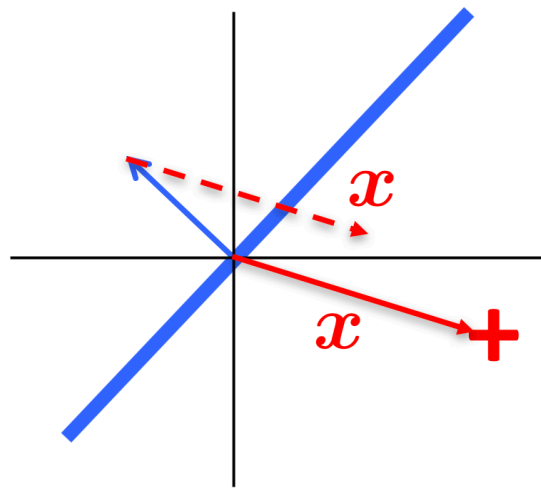
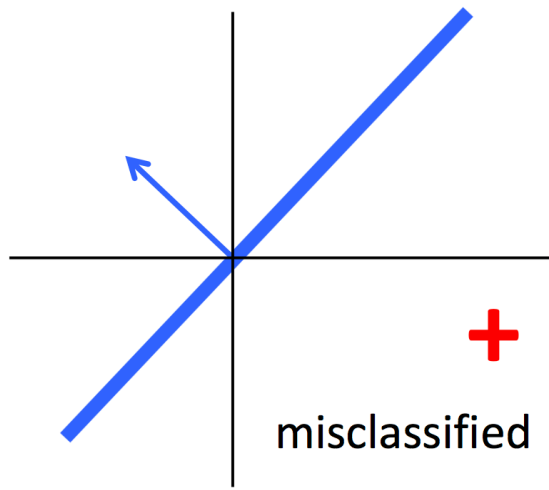
Cost function is 0 when classification is correct, and positive when incorrect

- 3) In the example to the right, how will the slope of the hyperplane change?

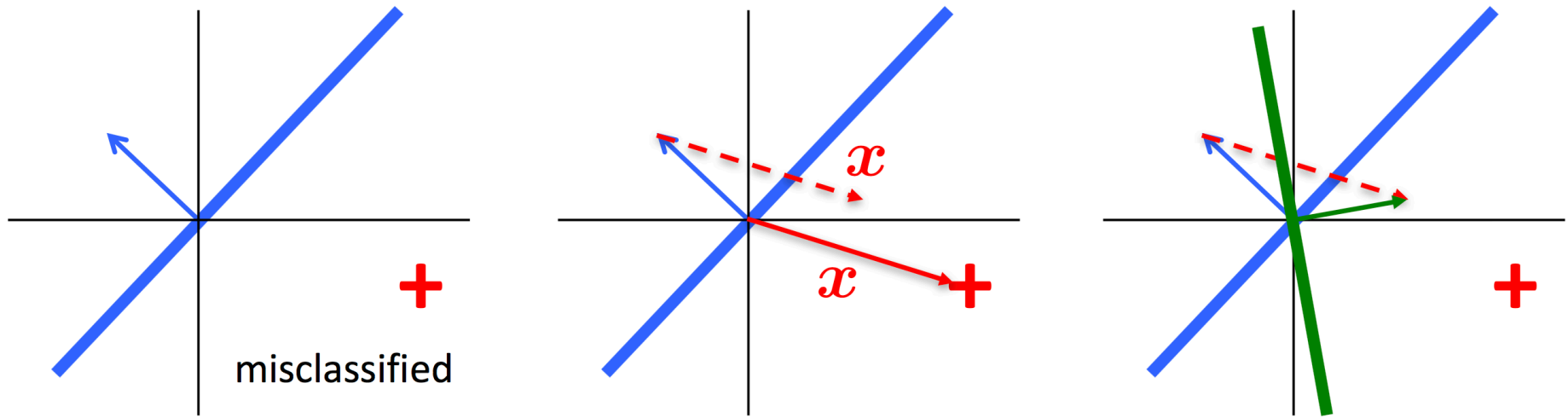


- 4) What are the weaknesses of the perceptron?
Create a binary classifier “wishlist”.

Perceptron algorithm and intuition



Perceptron algorithm and intuition



Let $\vec{w} = [0, 0, \dots, 0]^T$

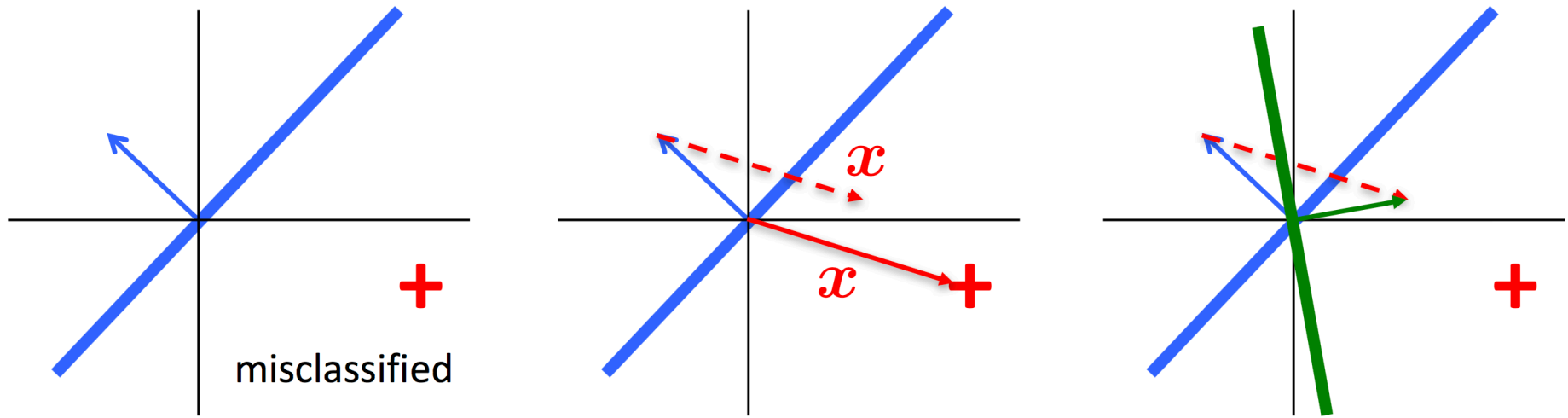
Repeat until convergence:

Receive training example (\vec{x}_i, y_i)

If $y_i(\vec{w}^T \vec{x}_i) \leq 0$ (incorrectly classified)

$$\vec{w} \leftarrow \vec{w} + \alpha y_i \vec{x}_i$$

Perceptron algorithm and intuition



Let $\vec{w} = [0, 0, \dots, 0]^T$

Repeat until convergence:

Receive training example (\vec{x}_i, y_i)

If $y_i(\vec{w}^T \vec{x}_i) \leq 0$ (incorrectly classified)

$$\vec{w} \leftarrow \vec{w} + \alpha y_i \vec{x}_i$$

Convergence:

- All data points correctly classified
- Fixed number of iterations passed

Often: $\alpha = 1$ (only changes magnitude of weight vector)

Binary classifier wishlist

- If data is linearly separable, want a “good” hyperplane (idea: far from points close to the boundary)
- If data is not linearly separable, want something reasonable (not just give up or fail to converge)
- Might not want to constrain ourselves to linear separators

Outline for March 22

- Perceptron history and interpretation as a neural network
- Idea of a maximum margin classifier
- Support Vector Machines introduction
- Functional vs. Geometric margins
- SVM as an optimization problem

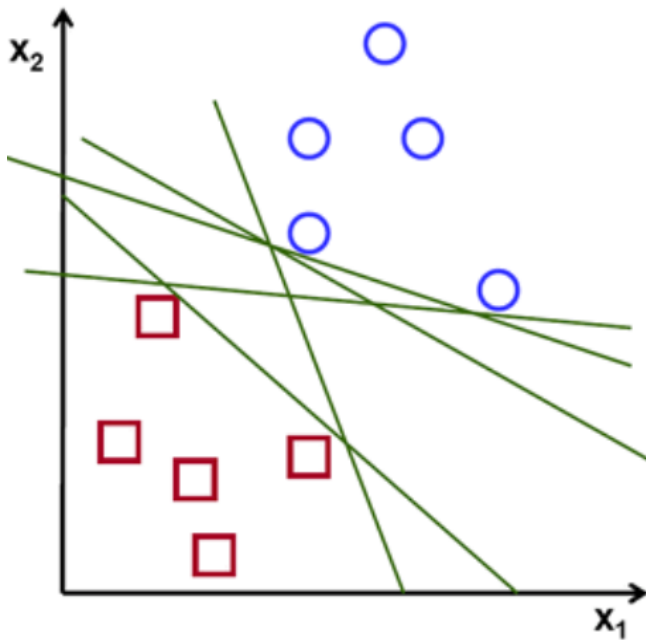
Support Vector Machines (SVMs)

- Will give us everything on our wishlist!
- Often considered the best “off the shelf” binary classifier
- Widely used in many fields

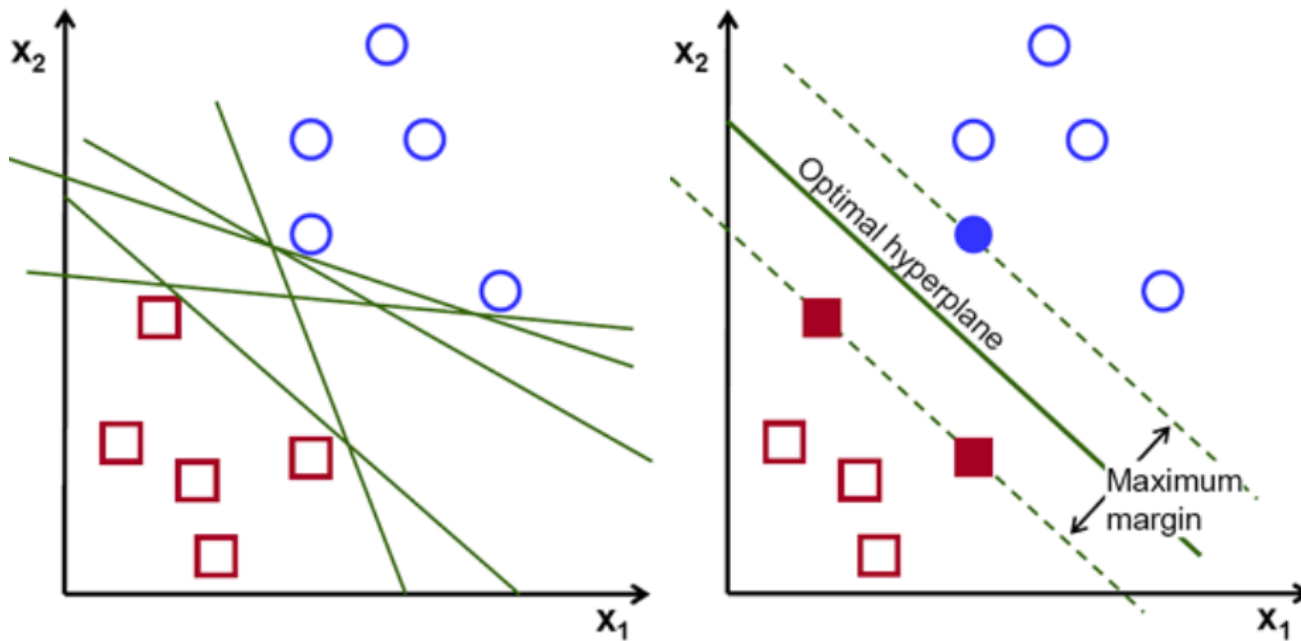
Brief history

- **1963**: Initial idea by Vladimir Vapnik and Alexey Chervonenkis
- **1992**: nonlinear SVMs by Bernhard Boser, Isabelle Guyon and Vladimir Vapnik
- **1993**: “soft-margin” by Corinna Cortes and Vladimir Vapnik

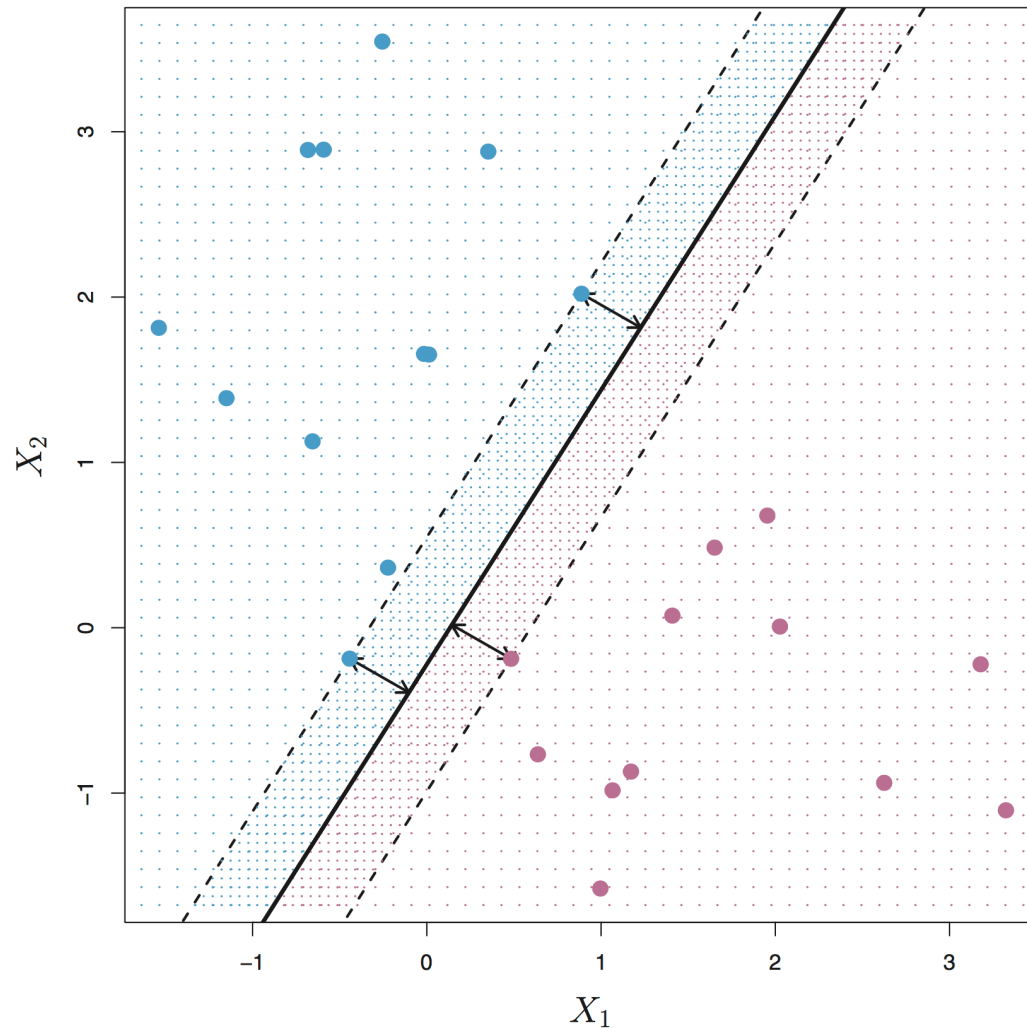
Idea: “best” hyperplane has a large margin



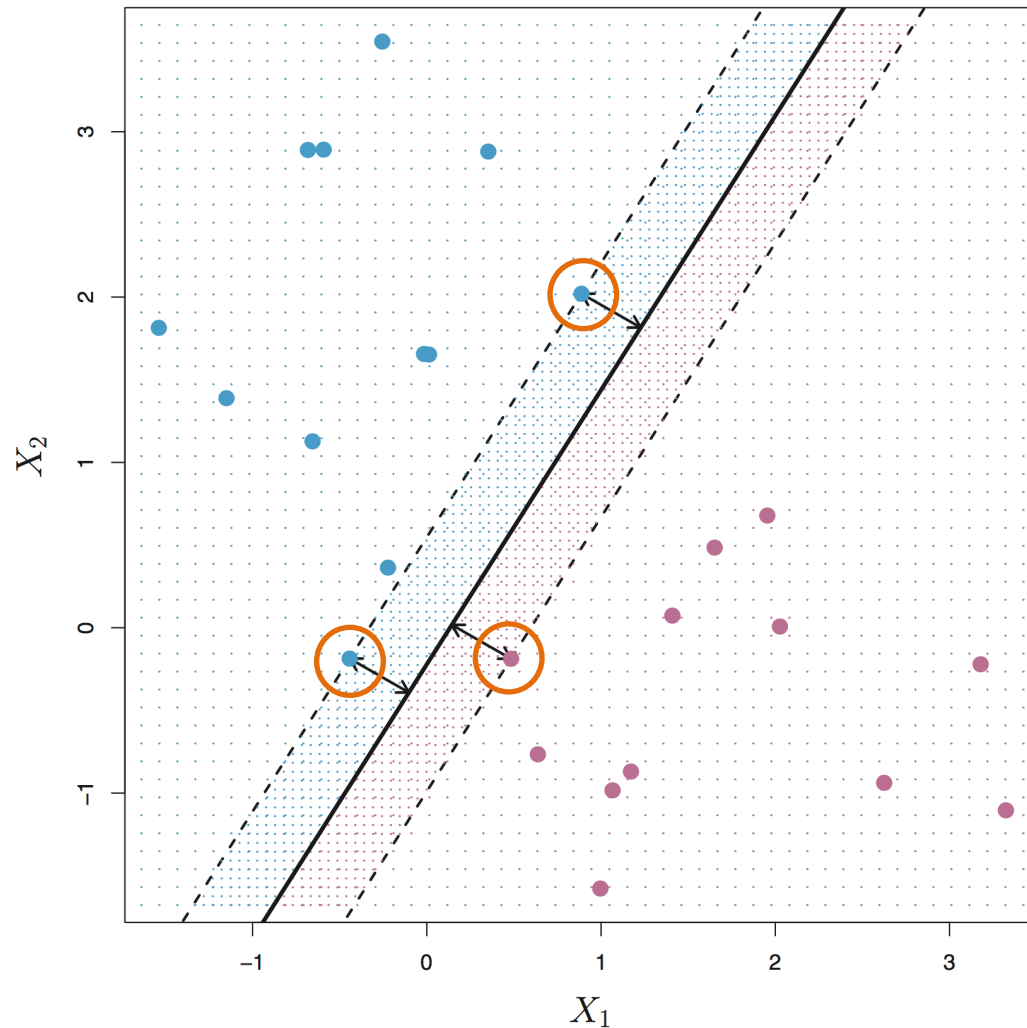
Idea: “best” hyperplane has a large margin



Datapoints that lie on the margin are called
“support vectors”



Datapoints that lie on the margin are called
“support vectors”



Support vectors

Outline for March 22

- Perceptron history and interpretation as a neural network
- Idea of a maximum margin classifier
- Support Vector Machines introduction
- Functional vs. Geometric margins
- SVM as an optimization problem

functional margin (-cost)

$$\hat{y}_i = y_i(\vec{w} \cdot \vec{x}_i + b)$$

if correct $\Rightarrow \hat{y}_i$ positive

else $\Rightarrow \hat{y}_i$ negative

bad: make $\vec{w} + b$ bigger
 \Rightarrow margin is bigger

good: can arbitrarily constrain

$\vec{w} + b$
(only care about
sign \hat{y}_i)

better: geometric margin

$$\vec{p} = \vec{x}_i - y_i \frac{\vec{w}}{\|\vec{w}\|}$$

on hyperplane!

which
direction
to subtract
off the
weight
vector

$$0 = \vec{w} \cdot \vec{p} + b$$

exercise

$$y_i = y_i \left(\frac{\vec{w}}{\|\vec{w}\|} \cdot \vec{x}_i + \frac{b}{\|\vec{w}\|} \right)$$

Goal: want to maximize

$$\min_{i=1, \dots, n} \gamma_i = \gamma$$

meaning:
w.r.t
these
variables

Optimization problem (try 1)

max	γ
s.t. (such that)	$\gamma_i(\vec{w} \cdot \vec{x}_i + b) \geq \gamma, i=1 \dots n$ $\ \vec{w}\ = 1$

} non-convex constraint.

try 2

Convex (one optimum)

Non-convex (many local optimum)

Example