CS182: Introduction to Machine Learning – Perceptron

Yujiao Shi

SIST, ShanghaiTech

Spring, 2025

Slides Courtesy of Matt Gormley & Henry Chai, CMU

THE PERCEPTRON ALGORITHM

Perceptron: History

上海科技大学 ShanghaiTech University

Imagine you are trying to build a new machine learning technique… your name is Frank Rosenblatt…and the year is 1957

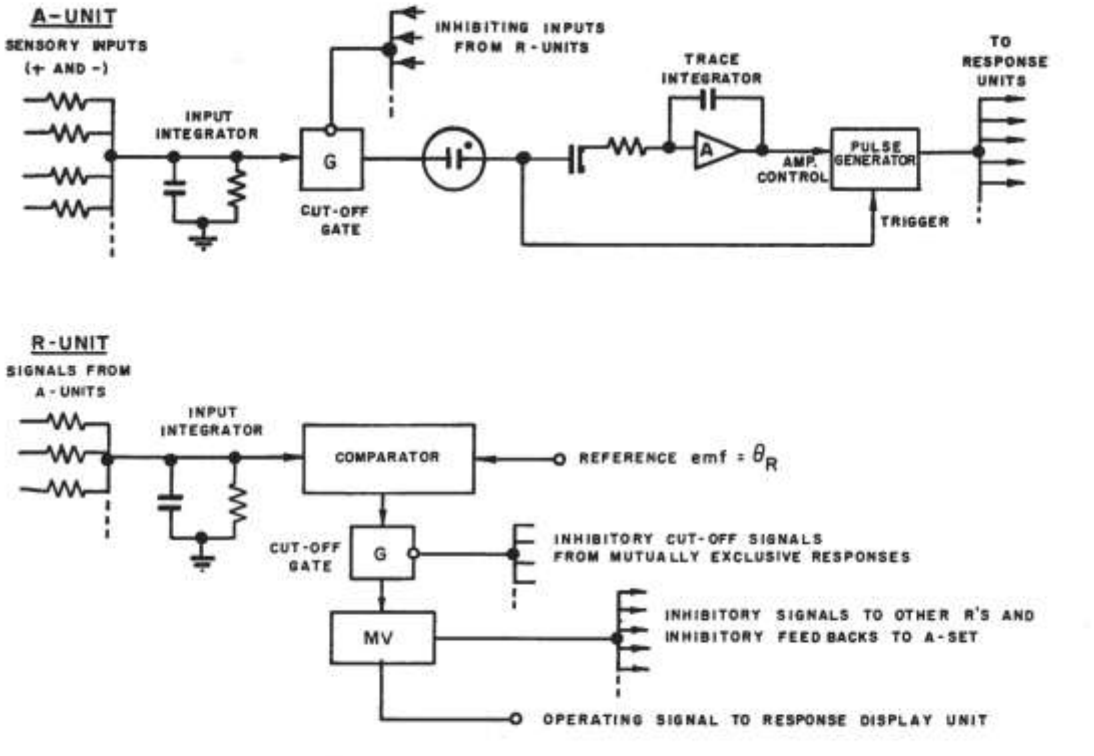
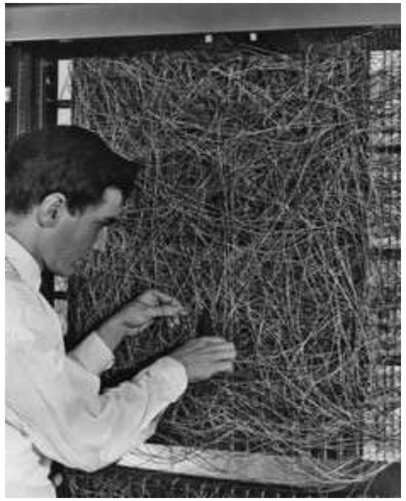


FIGURE 5
DESIGN OF TYPICAL UNITS

Perceptron: History

上海科技大学 ShanghaiTech University

Imagine you are trying to build a new machine learning technique… your name is Frank Rosenblatt… and the year is 1957

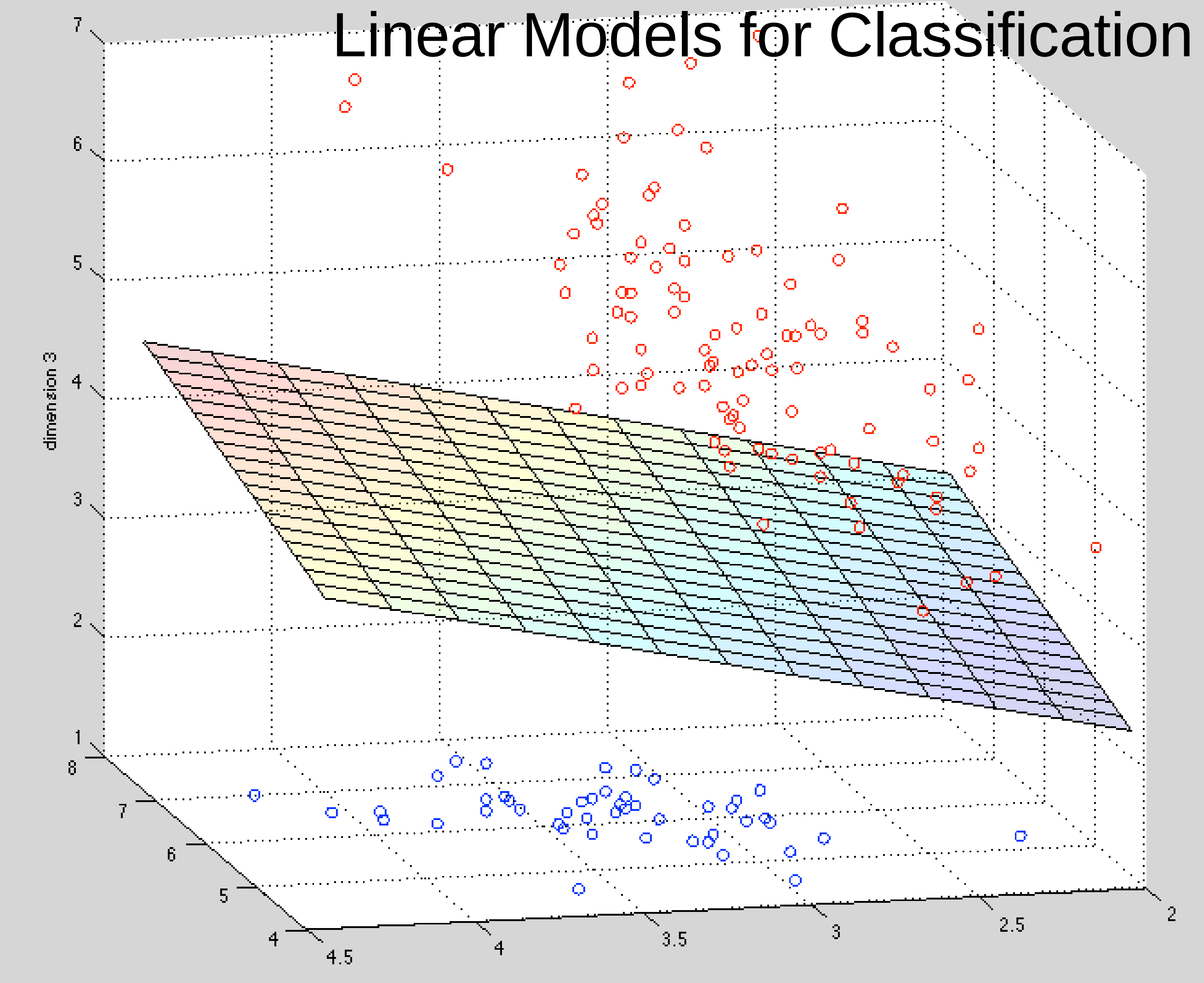


The New Yorker, December 6, 1958 P. 44

Talk story about the perceptron, a new electronic brain which hasn't been built, but which has been successfully simulated on the I.B.M. 704. Talk with Dr. Frank Rosenblatt, of the Cornell Aeronautical Laboratory, who is one of the two men who developed the prodigy; the other man is Dr. Marshall C. Yovits, of the Office of Naval Research, in Washington. Dr. Rosenblatt defined the perceptron as the first non-biological object which will achieve an organization o its external environment in a meaningful way. It interacts with its environment, forming concepts that have not been made ready for it by a human agent. If a triangle is held up, the perceptron's eye picks up the image & conveys it along a random succession of lines to the response units, where the image is registered. It can tell the difference betw. a cat and a dog, although it wouldn't be able to tell whether the dog was to theleft or right of the cat. Right now it is of no practical use, Dr. Rosenblatt conceded, but he said that one day it might be useful to send one into outer space to take in impressions for us.

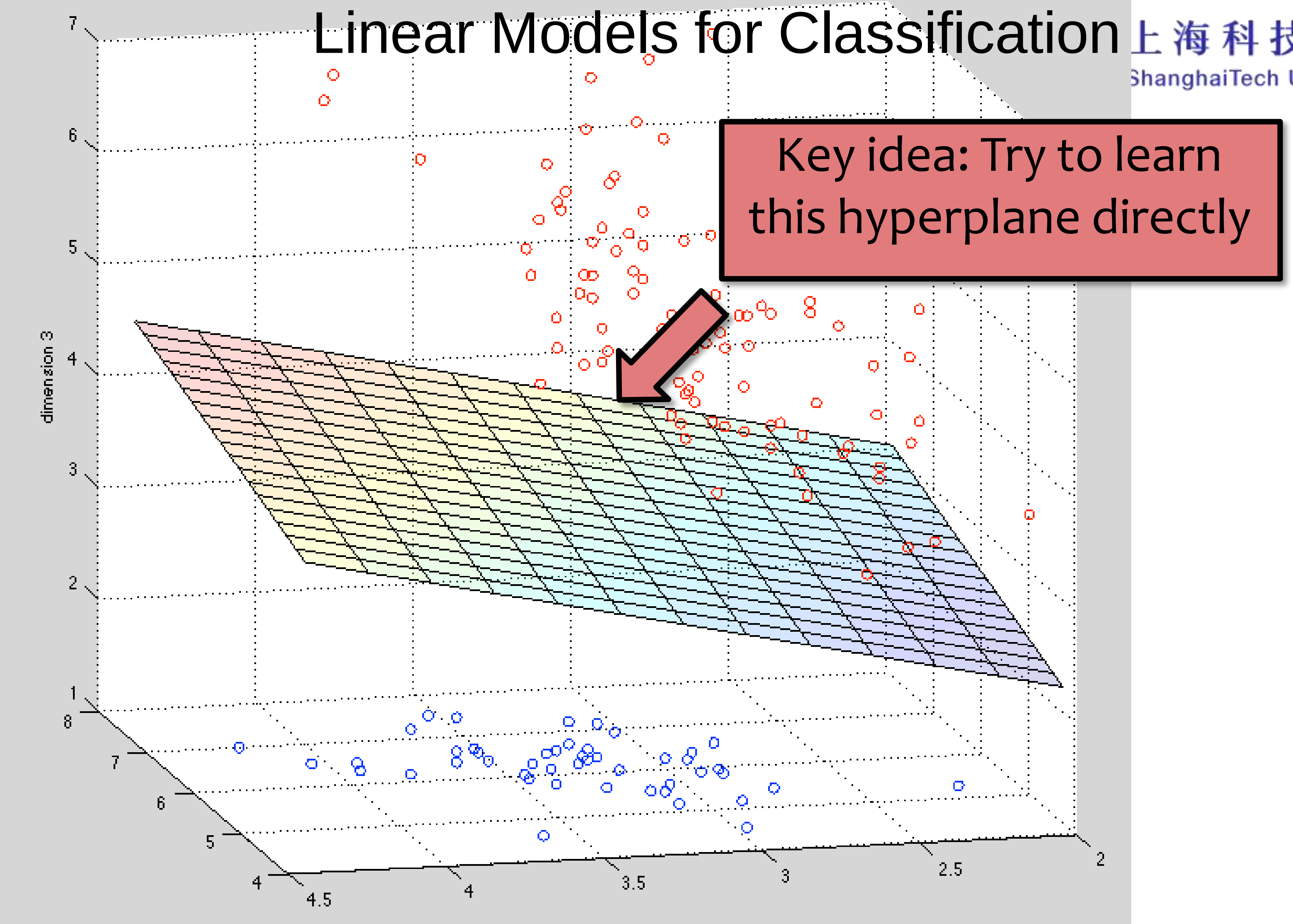
Linear Models for Classification

上海科技大学 ShanghaiTech University

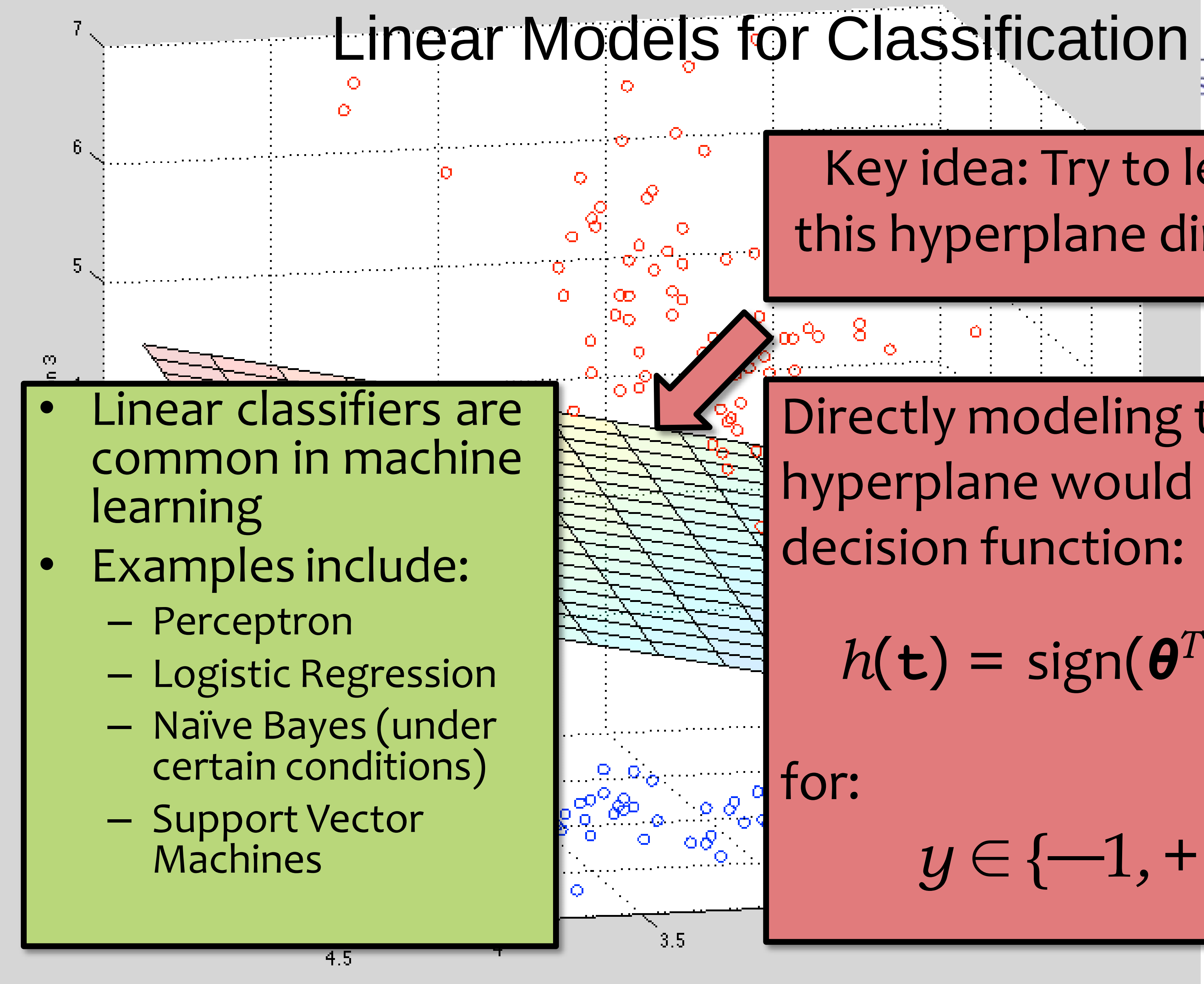


Linear Models for Classification

上海科技大学 ShanghaiTech University



Linear Models for Classification



- Linear classifiers are common in machine learning
- Examples include:
  - Perceptron
  - Logistic Regression
  - Naïve Bayes (under certain conditions)
  - Support Vector Machines

Key idea: Try to learn this hyperplane directly

Directly modeling the hyperplane would use a decision function:

$$h(\mathbf{t}) = \text{sign}(\boldsymbol{\theta}^T \mathbf{t})$$

for:

$$y \in \{-1, +1\}$$

上海科技大学
ShanghaiTech University

GEOMETRY & VECTORS

In-Class Exercise

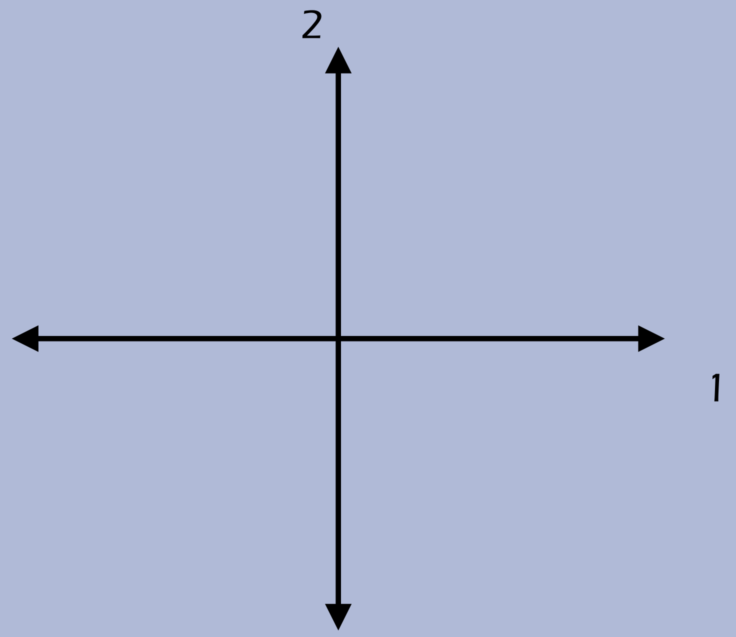
Draw a picture of the region corresponding to:

$$w_1 x_1 + w_2 x_2 + b > 0$$

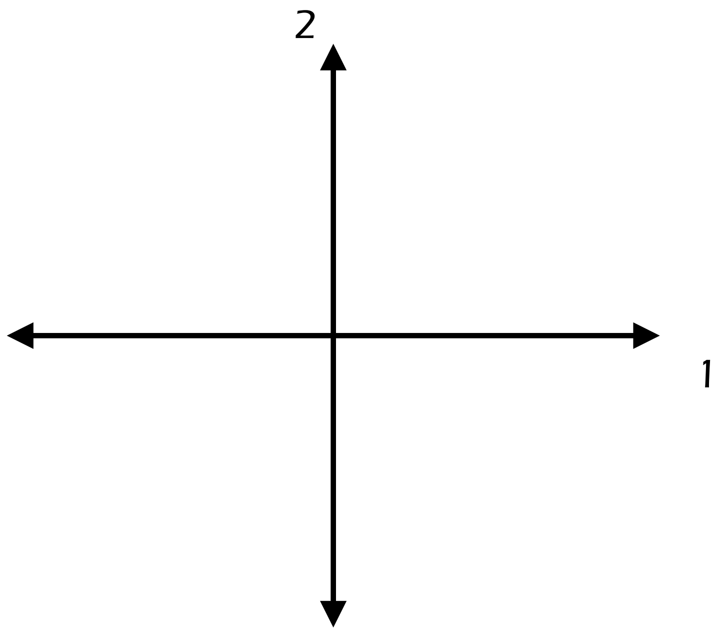
where $w_1 = 2, w_2 = 3, b = 6$

Draw the vector **w** = $[w_1, w_2]$

Answer Here:



Geometry Warm-up



- Notation: in this class vectors will be assumed to be column vectors by default, i.e.,

$$\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_D \end{bmatrix} \text{ and } \boldsymbol{a}^T = [a_1 \quad a_2 \quad \cdots \quad a_D]$$

- The dot product between two D-dimensional vectors is

$$\boldsymbol{a}^T\boldsymbol{b} = [a_1 \quad a_2 \quad \cdots \quad a_D] \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_D \end{bmatrix} = \sum_{d=1}^{D} a_d b_d$$

- The $L2$-norm of $\boldsymbol{a} = \|\boldsymbol{a}\|_2 = \sqrt{\boldsymbol{a}^T\boldsymbol{a}}$
- Two vectors are *orthogonal* iff

$$\boldsymbol{a}^T\boldsymbol{b} = 0$$

Vector Projection

Question:

Which of the following is the projection of a vector ***a*** *onto a vector* ***b****?*

A. $\frac{\mathbf{a}^T\mathbf{b}}{\mathbf{b}}\mathbf{a}$

B. $\frac{\mathbf{a}\cdot\mathbf{b}}{\mathbf{a}^T\mathbf{b}}$

C. $\frac{(\mathbf{a}^T\mathbf{b})}{||\mathbf{b}||_2}\mathbf{b}$

D. $\frac{(\mathbf{a}\cdot\mathbf{b})}{||\mathbf{b}||_2}\mathbf{b}$

E. $\frac{(\mathbf{a}^T\mathbf{b})}{||\mathbf{b}||_2^2}\mathbf{b}$

F. $\frac{(\mathbf{a}^T\mathbf{b})^2}{||\mathbf{b}||_2}\mathbf{b}$

Vector Projection

Definition #1:

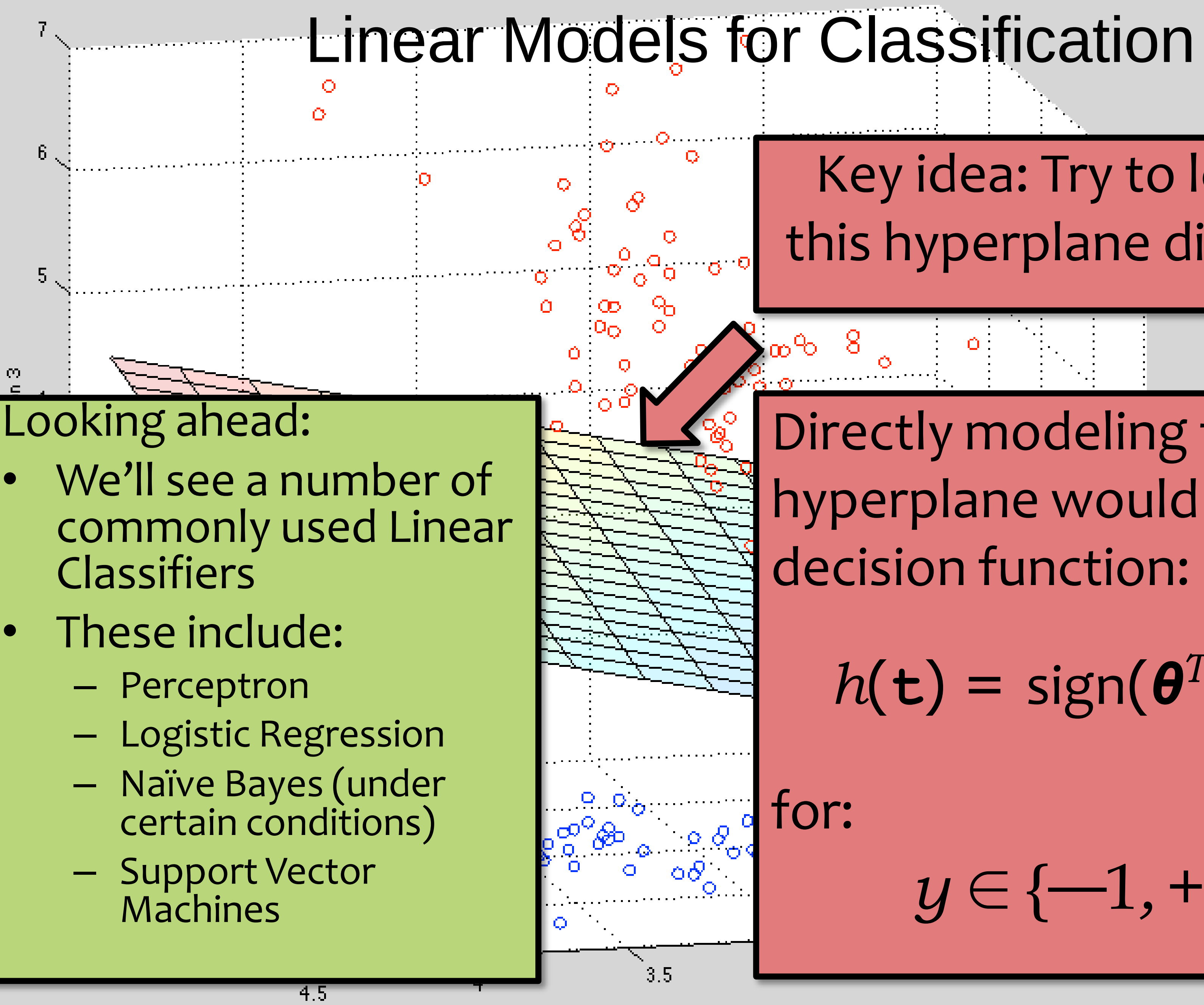
Definition #2:

Linear Decision Boundaries

上海科技大学 ShanghaiTech University

- In 2 dimensions, $w_1 x_1 + w_2 x_2 + b = 0$ defines a *line*
- In 3 dimensions, $w_1 x_1 + w_2 x_2 + w_3 x_3 + b = 0$ defines a *plane*
- In 4+ dimensions, $\boldsymbol{w}^T \boldsymbol{x} + b = 0$ defines a *hyperplane*
  - The vector $\boldsymbol{w}$ is always orthogonal to this hyperplane and always points in the direction where $\boldsymbol{w}^T \boldsymbol{x} + b > 0$!
- A hyperplane creates two *halfspaces*:
  - $\mathrm{S}_+ = \{\boldsymbol{x}:\ \boldsymbol{w}^T \boldsymbol{x} + b > 0\}$ or all $\boldsymbol{x}$ s.t. $\boldsymbol{w}^T \boldsymbol{x} + b$ is positive
  - $\mathrm{S}_- = \{\boldsymbol{x}:\ \boldsymbol{w}^T \boldsymbol{x} + b < 0\}$ or all $\boldsymbol{x}$ s.t. $\boldsymbol{w}^T \boldsymbol{x} + b$ is negative

Linear Models for Classification



Key idea: Try to learn this hyperplane directly

Looking ahead:

- We'll see a number of commonly used Linear Classifiers
- These include:
  - Perceptron
  - Logistic Regression
  - Naïve Bayes (under certain conditions)
  - Support Vector Machines

Directly modeling the hyperplane would use a decision function:

$$h(\mathbf{t}) = \text{sign}(\boldsymbol{\theta}^T \mathbf{t})$$

for:

$$y \in \{-1, +1\}$$

ONLINE LEARNING

Online Learning

- **Batch Learning:** So far, we've been learning in the *batch* setting, where we have access to the entire training dataset at once
- **Online Learning:** A common alternative is the *online* setting, where examples arrive gradually and we learn continuously

- **Examples of online learning:**
  1. **Stock market** prediction (what will the value of Alphabet Inc. be tomorrow?)
  2. **Email** classification (distribution of both spam and regular mail changes over time, but the target function stays fixed - last year's spam still looks like spam)
  3. **Recommendation** systems. Examples: recommending movies; predicting whether a user will be interested in a new news article
  4. **Ad placement** in a new market

Slide adapted from Nina Balcan

For i = 1, 2, 3, … **:**

– **Receive** an unlabeled instance $\mathbf{x}^{(i)}$

– **Predict** $y' = h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})$

– **Receive** true label $y^{(i)}$

– **Suffer loss** if we made a mistake, $y' \neq y^{(i)}$

– **Update** parameters $\boldsymbol{\theta}$

Goal:

– **Minimize** the number of **mistakes**

THE PERCEPTRON ALGORITHM

(Online) Perceptron Algorithm

上海科技大学 ShanghaiTech University

- Initialize the weight vector and intercept to all zeros:

  $\boldsymbol{w} = [0 \quad 0 \quad \cdots \quad 0]$ and $b = 0$

- For $t = 1, 2, 3, \ldots$
  - Receive an unlabeled example, $\boldsymbol{x}^{(t)}$
  - Predict its label, $\hat{y} = \text{sign}\,(\boldsymbol{w}^T\boldsymbol{x} + b) = \begin{cases} +1 \text{ if } \boldsymbol{w}^T\boldsymbol{x} + b \geq 0 \\ -1 \text{ otherwise} \end{cases}$
  - Observe its true label, $y^{(t)}$
  - If we misclassified a positive example ($y^{(t)} = +1, \hat{y} = -1$):
    - $\boldsymbol{w} \leftarrow \boldsymbol{w} + \boldsymbol{x}^{(t)}$
    - $b \leftarrow b + 1$
  - If we misclassified a negative example ($y^{(t)} = -1, \hat{y} = +1$):
    - $\boldsymbol{w} \leftarrow \boldsymbol{w} - \boldsymbol{x}^{(t)}$
    - $b \leftarrow b - 1$

Learning Intuition:

Learning Intuition:

- Suppose $(\boldsymbol{x}, y) \in \mathcal{D}$ is a misclassified training example and $y = +1$
  - $\boldsymbol{\theta}^T \boldsymbol{x}$ is negative
  - After updating $\boldsymbol{\theta}_{new} = \boldsymbol{\theta} + y\boldsymbol{x}$:

$$\boldsymbol{\theta}_{new}^T \boldsymbol{x} = (\boldsymbol{\theta} + y\boldsymbol{x})^T \boldsymbol{x} = \boldsymbol{\theta}^T \boldsymbol{x} + y\boldsymbol{x}^T \boldsymbol{x}$$

  which is less negative than $\boldsymbol{\theta}^T \boldsymbol{x}$
    - Because $y > 0$ and $\boldsymbol{x}^T \boldsymbol{x} > 0$
  - Our prediction for $\boldsymbol{x}$ “improved”!
- A similar argument holds if $y = -1$

(Online) Perceptron Algorithm

上海科技大学 ShanghaiTech University

- Initialize the weight vector and intercept to all zeros:

  $\boldsymbol{w} = [0 \quad 0 \quad \cdots \quad 0]$ and $b = 0$

- For $t = 1, 2, 3, \ldots$
  - Receive an unlabeled example, $\boldsymbol{x}^{(t)}$
  - Predict its label, $\hat{y} = \text{sign}\,(\boldsymbol{w}^T\boldsymbol{x} + b) = \begin{cases} +1 \text{ if } \boldsymbol{w}^T\boldsymbol{x} + b \geq 0 \\ -1 \text{ otherwise} \end{cases}$
  - Observe its true label, $y^{(t)}$
  - If we misclassified a positive example ($y^{(t)} = +1, \hat{y} = -1$):
    - $\boldsymbol{w} \leftarrow \boldsymbol{w} + \boldsymbol{x}^{(t)}$
    - $b \leftarrow b + 1$
  - If we misclassified a negative example ($y^{(t)} = -1, \hat{y} = +1$):
    - $\boldsymbol{w} \leftarrow \boldsymbol{w} - \boldsymbol{x}^{(t)}$
    - $b \leftarrow b - 1$

(Online) Perceptron Algorithm

上海科技大学 ShanghaiTech University

- Initialize the weight vector and intercept to all zeros:

  $\boldsymbol{w} = [0 \quad 0 \quad \cdots \quad 0]$ and $b = 0$

- For $t = 1, 2, 3, \ldots$
  - Receive an unlabeled example, $\boldsymbol{x}^{(t)}$
  - Predict its label, $\hat{y} = \text{sign}\,(\boldsymbol{w}^T\boldsymbol{x} + b) = \begin{cases} +1 \text{ if } \boldsymbol{w}^T\boldsymbol{x} + b \geq 0 \\ -1 \text{ otherwise} \end{cases}$
  - Observe its true label, $y^{(t)}$
  - If we misclassified an example ($y^{(t)} \neq \hat{y}$):
    - $\boldsymbol{w} \leftarrow \boldsymbol{w} + y^{(t)}\boldsymbol{x}^{(t)}$
    - $b \leftarrow b + y^{(t)}$

Implementation trick: Multiplying by y^{t} gives us a simple update rule for both positive *and* negative mistakes

The **Great Elephant Census** of 2014 revealed that elephant populations were **trending downward** at an alarming rate.

Poaching is known to be one of the main threats to elephants.

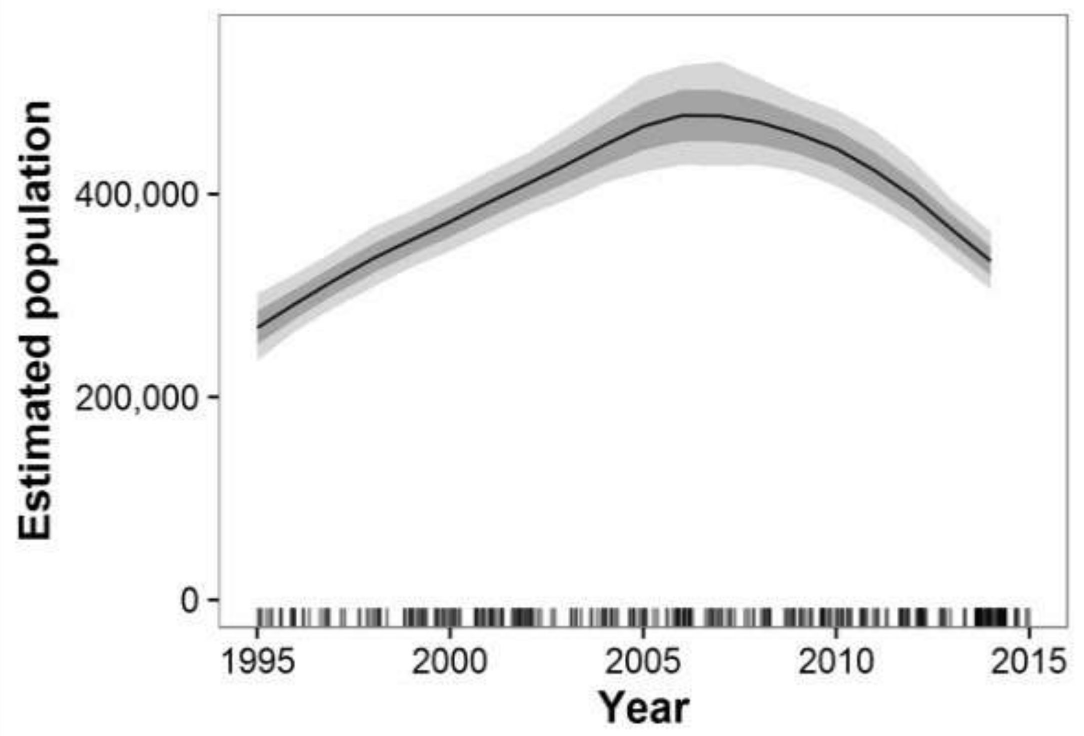


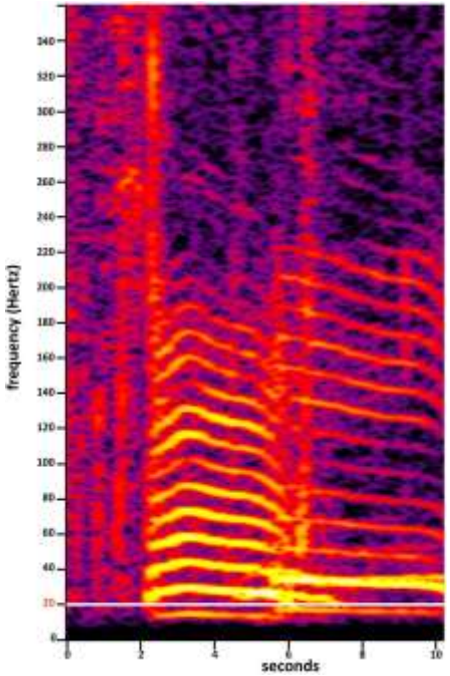
Figure 2: Estimated trends in elephant populations for GEC study areas with historical data available, 1995–2014.
Results are based on 1,000 Monte Carlo replicates. Dark shaded area indicates ±1 SD; light shaded area indicates 95% confidence interval. Tick marks on *x*-axis indicate dates of data points used in model; dates are perturbed slightly to prevent overlap.

Download full-size image

DOI: 10.7717/peerj.2354/fig-2

https://peerj.com/articles/2354/

AI for Wildlife Conservation



- Researchers at Cornell planted **50 audio recording devices** high in the jungle -- each one covering a 25 square km grid cell
- Recordings revealed two large creatures making noise: **elephants** and **poachers**
- So they built **classifiers** to detect these

https://www.npr.org/2019/10/25/760487476/elephants-under-attack-have-an-unlikely-ally-artificial-intelligence

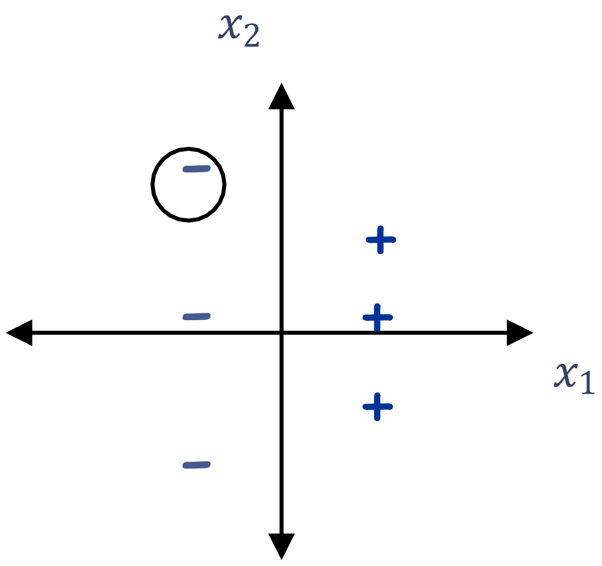
(Online) Perceptron Algorithm: Example

Perceptron Algorithm: (without the intercept term)

- Set t=1, start with all-zeroes weight vector w_1.
- Given example x, predict positive iff $w_t \cdot x \geq 0$.
- On a mistake, update as follows:
  - Mistake on positive, update $w_{t+1} \leftarrow w_t + x$
  - Mistake on negative, update $w_{t+1} \leftarrow w_t - x$

x_1	x_2	$\hat{y}$	y	Mistake?
-1	2	$+$	$-$	Yes

$$\boldsymbol{w} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$



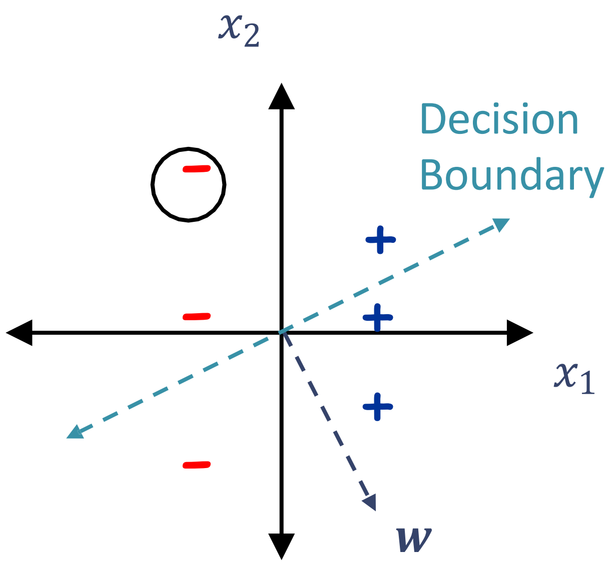
Example courtesy of Nina Balcan

(Online) Perceptron Algorithm: Example

Perceptron Algorithm: (without the intercept term)

- Set t=1, start with all-zeroes weight vector w_1.
- Given example x, predict positive iff $w_t \cdot x \geq 0$.
- On a mistake, update as follows:
  - Mistake on positive, update $w_{t+1} \leftarrow w_t + x$
  - Mistake on negative, update $w_{t+1} \leftarrow w_t - x$

x_1	x_2	$\hat{y}$	y	Mistake?
-1	2	+	−	Yes



$$\boldsymbol{w} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + y^{(1)}\boldsymbol{x}^{(1)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

Example courtesy of Nina Balcan

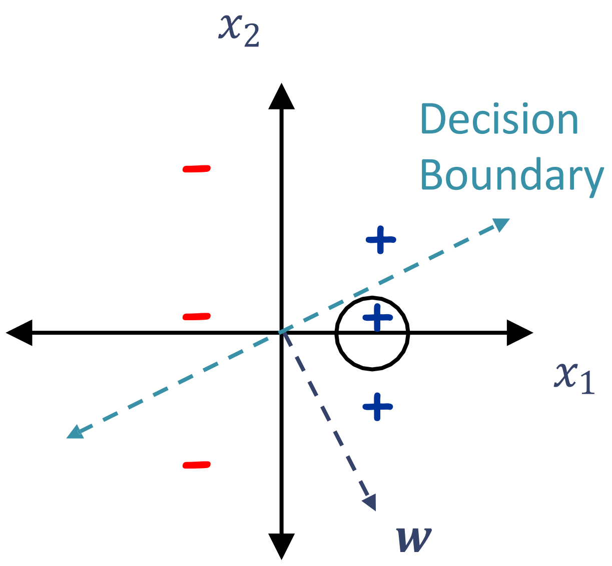
(Online) Perceptron Algorithm: Example

Perceptron Algorithm: (without the intercept term)

- Set t=1, start with all-zeroes weight vector w_1.
- Given example x, predict positive iff $w_t \cdot x \geq 0$.
- On a mistake, update as follows:
  - Mistake on positive, update $w_{t+1} \leftarrow w_t + x$
  - Mistake on negative, update $w_{t+1} \leftarrow w_t - x$

x_1	x_2	$\hat{y}$	y	Mistake?
-1	2	$+$	$-$	Yes
1	0	$+$	$+$	No

$$\boldsymbol{w} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$



Example courtesy of Nina Balcan

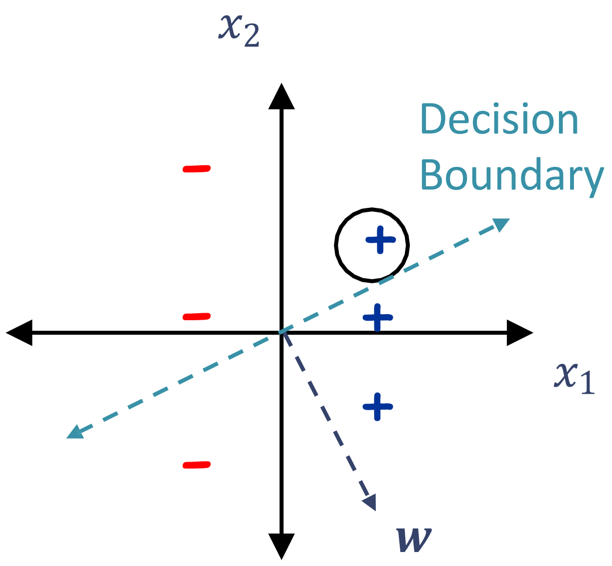
(Online) Perceptron Algorithm: Example

上海科技大学 ShanghaiTech University

Perceptron Algorithm: (without the intercept term)

- Set t=1, start with all-zeroes weight vector w_1.
- Given example x, predict positive iff $w_t \cdot x \geq 0$.
- On a mistake, update as follows:
  - Mistake on positive, update $w_{t+1} \leftarrow w_t + x$
  - Mistake on negative, update $w_{t+1} \leftarrow w_t - x$

x_1	x_2	$\hat{y}$	y	Mistake?
-1	2	$+$	$-$	Yes
1	0	$+$	$+$	No
1	1	$-$	$+$	Yes



$$\boldsymbol{w} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + y^{(3)}\boldsymbol{x}^{(3)} = \begin{bmatrix} 1 \\ -2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

Example courtesy of Nina Balcan

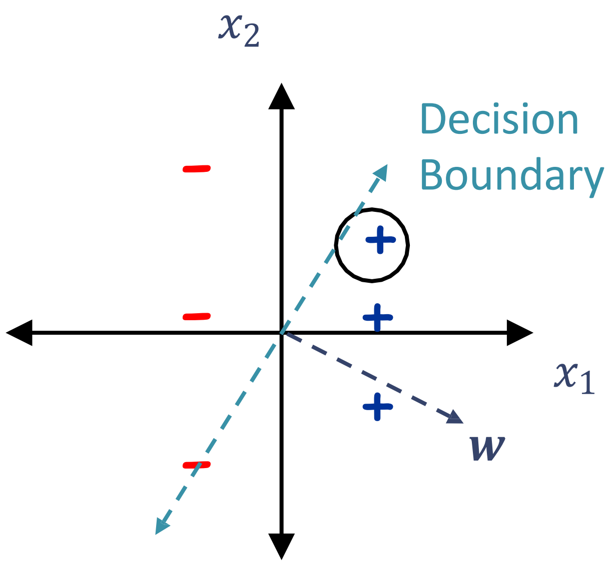
(Online) Perceptron Algorithm: Example

上海科技大学
ShanghaiTech University

Perceptron Algorithm: (without the intercept term)

- Set t=1, start with all-zeroes weight vector w_1.
- Given example x, predict positive iff $w_t \cdot x \geq 0$.
- On a mistake, update as follows:
  - Mistake on positive, update $w_{t+1} \leftarrow w_t + x$
  - Mistake on negative, update $w_{t+1} \leftarrow w_t - x$

x_1	x_2	$\hat{y}$	y	Mistake?
-1	2	$+$	$-$	Yes
1	0	$+$	$+$	No
1	1	$-$	$+$	Yes



$$\boldsymbol{w} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + y^{(3)}\boldsymbol{x}^{(3)} = \begin{bmatrix} 1 \\ -2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

Example courtesy of Nina Balcan

(Online) Perceptron Algorithm: Example

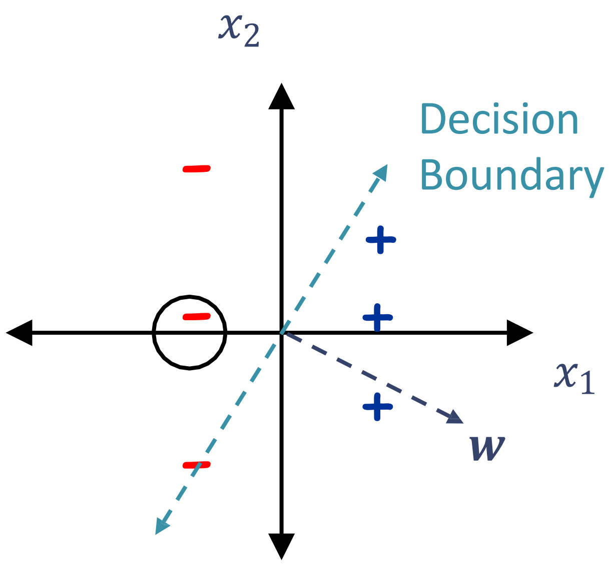
上海科技大学 ShanghaiTech University

Perceptron Algorithm: (without the intercept term)

- Set t=1, start with all-zeroes weight vector w_1.
- Given example x, predict positive iff $w_t \cdot x \geq 0$.
- On a mistake, update as follows:
  - Mistake on positive, update $w_{t+1} \leftarrow w_t + x$
  - Mistake on negative, update $w_{t+1} \leftarrow w_t - x$

x_1	x_2	$\hat{y}$	y	Mistake?
-1	2	$+$	$-$	Yes
1	0	$+$	$+$	No
1	1	$-$	$+$	Yes
-1	0	$-$	$-$	No

$$\boldsymbol{w} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$



Example courtesy of Nina Balcan

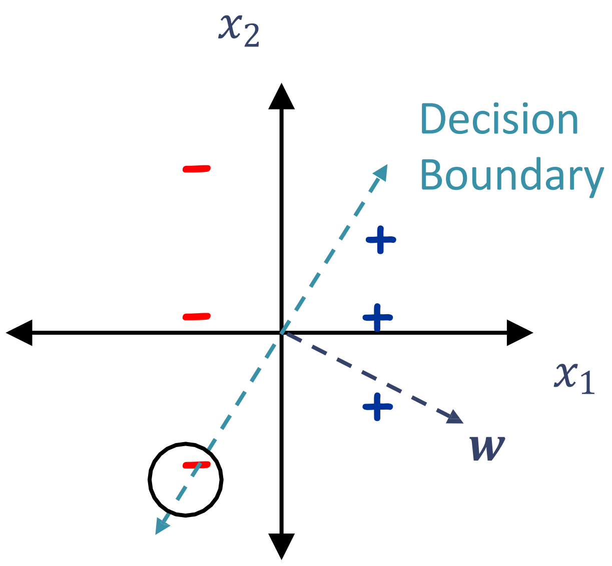
(Online) Perceptron Algorithm: Example

上海科技大学 ShanghaiTech University

Perceptron Algorithm: (without the intercept term)

- Set t=1, start with all-zeroes weight vector w_1.
- Given example x, predict positive iff $w_t \cdot x \geq 0$.
- On a mistake, update as follows:
  - Mistake on positive, update $w_{t+1} \leftarrow w_t + x$
  - Mistake on negative, update $w_{t+1} \leftarrow w_t - x$

x_1	x_2	$\hat{y}$	y	Mistake?
-1	2	$+$	$-$	Yes
1	0	$+$	$+$	No
1	1	$-$	$+$	Yes
-1	0	$-$	$-$	No
-1	-2	$+$	$-$	Yes



$$\boldsymbol{w} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + y^{(5)}\boldsymbol{x}^{(5)} = \begin{bmatrix} 2 \\ -1 \end{bmatrix} - \begin{bmatrix} -1 \\ -2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

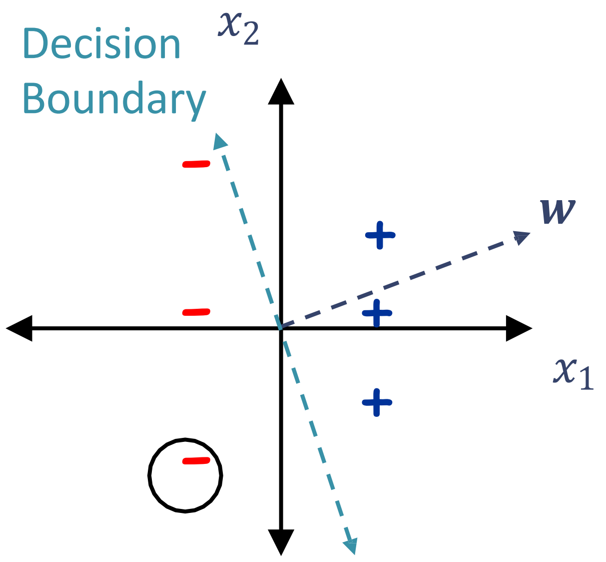
Example courtesy of Nina Balcan

(Online) Perceptron Algorithm: Example

Perceptron Algorithm: (without the intercept term)

- Set t=1, start with all-zeroes weight vector w_1.
- Given example x, predict positive iff $w_t \cdot x \geq 0$.
- On a mistake, update as follows:
  - Mistake on positive, update $w_{t+1} \leftarrow w_t + x$
  - Mistake on negative, update $w_{t+1} \leftarrow w_t - x$

x_1	x_2	$\hat{y}$	y	Mistake?
-1	2	$+$	$-$	Yes
1	0	$+$	$+$	No
1	1	$-$	$+$	Yes
-1	0	$-$	$-$	No
-1	-2	$+$	$-$	Yes



$$\boldsymbol{w} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + y^{(5)}\boldsymbol{x}^{(5)} = \begin{bmatrix} 2 \\ -1 \end{bmatrix} - \begin{bmatrix} -1 \\ -2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

Example courtesy of Nina Balcan

(Online) Perceptron Algorithm: Example

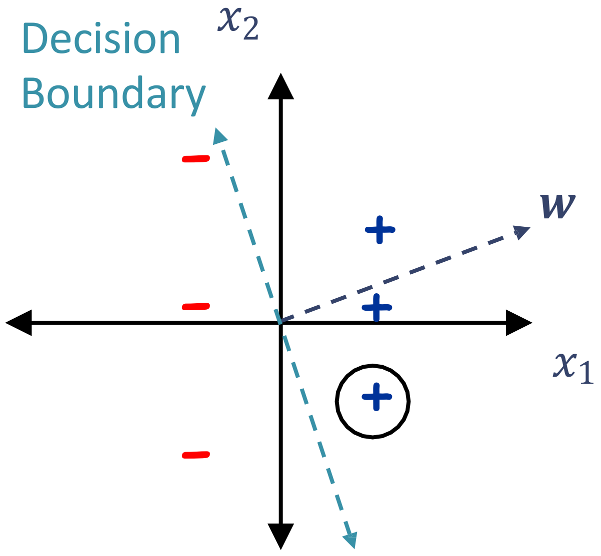
上海科技大学 ShanghaiTech University

Perceptron Algorithm: (without the intercept term)

- Set t=1, start with all-zeroes weight vector w_1.
- Given example x, predict positive iff $w_t \cdot x \geq 0$.
- On a mistake, update as follows:
  - Mistake on positive, update $w_{t+1} \leftarrow w_t + x$
  - Mistake on negative, update $w_{t+1} \leftarrow w_t - x$

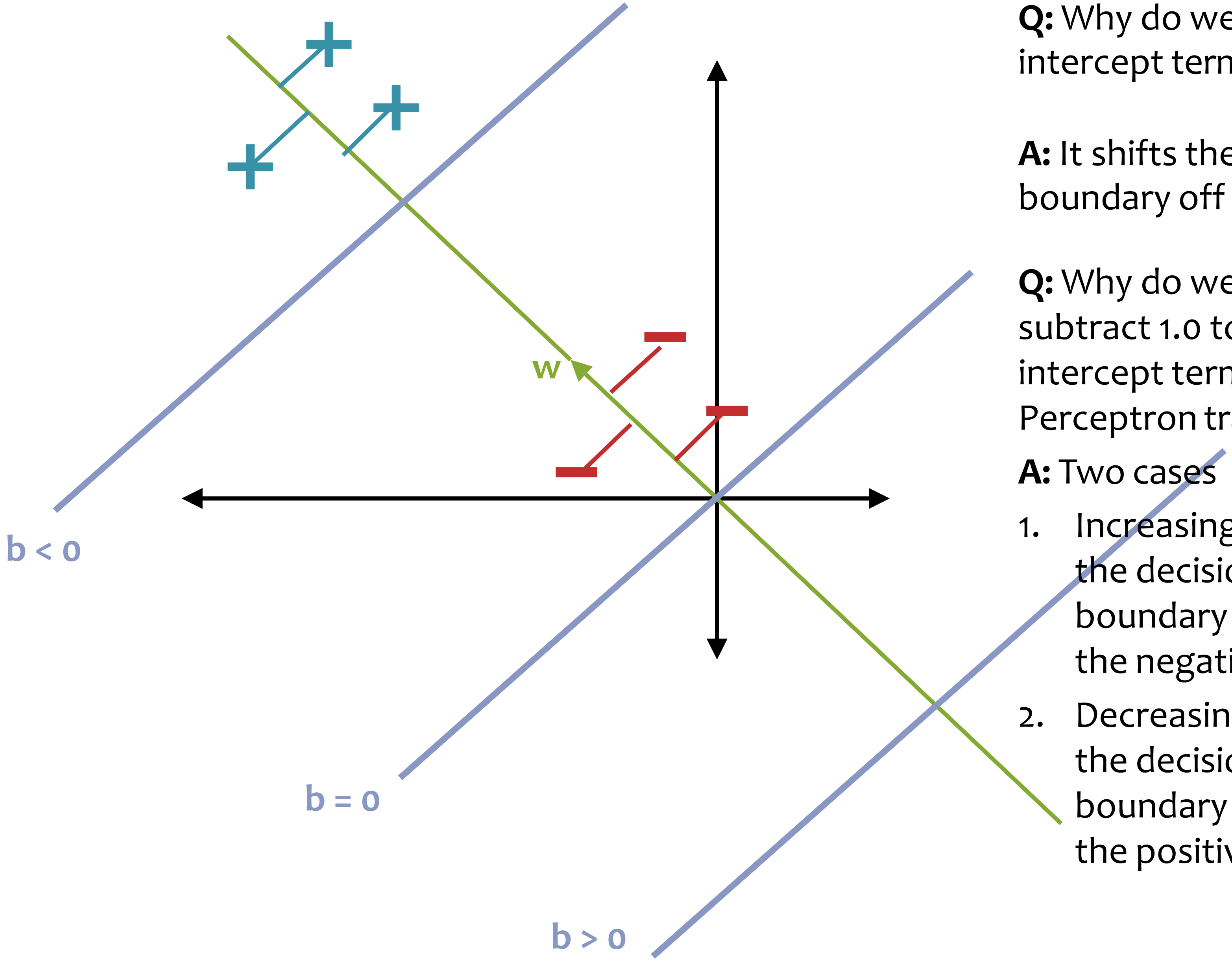
x_1	x_2	$\hat{y}$	y	Mistake?
-1	2	$+$	$-$	Yes
1	0	$+$	$+$	No
1	1	$-$	$+$	Yes
-1	0	$-$	$-$	No
-1	-2	$+$	$-$	Yes
1	-1	$+$	$+$	No

$$\boldsymbol{w} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$



Example courtesy of Nina Balcan

Intercept Term



Q: Why do we need an intercept term?

A: It shifts the decision boundary off the origin

Q: Why do we add / subtract 1.0 to the intercept term during Perceptron training?

A: Two cases

1. Increasing b shifts the decision boundary towards the negative side
2. Decreasing b shifts the decision boundary towards the positive side

Question:

The parameter vector **w** *learned by the Perceptron algorithm can be* ***written as a linear combination*** *of the feature vectors* $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}$.

A. *True, if you replace "linear" with "polynomial" above*
B. *True, for all datasets*
C. *False, for all datasets*
D. *True, but only for certain datasets*
E. *False, but only for certain datasets*

Notational Hack

- If we add a 1 to the beginning of every example e.g.,

$$\boldsymbol{x}' = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} \ldots$$

- … we can just fold the intercept into the weight vector!

$$\boldsymbol{\theta} = \begin{bmatrix} b \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix} \rightarrow \boldsymbol{\theta}^T \boldsymbol{x}' = \boldsymbol{w}^T \boldsymbol{x} + b$$

(Online) Perceptron Algorithm

上海科技大学 ShanghaiTech University

- Initialize the weight vector and intercept to all zeros:

  $\boldsymbol{w} = [0 \quad 0 \quad \cdots \quad 0]$ and $b = 0$

- For $t = 1, 2, 3, \ldots$
  - Receive an unlabeled example, $\boldsymbol{x}^{(t)}$
  - Predict its label, $\hat{y} = \text{sign}\,(\boldsymbol{w}^T\boldsymbol{x} + b) = \begin{cases} +1 \text{ if } \boldsymbol{w}^T\boldsymbol{x} + b \geq 0 \\ -1 \text{ otherwise} \end{cases}$
  - Observe its true label, $y^{(t)}$
  - If we misclassified an example ($y^{(t)} \neq \hat{y}$):
    - $\boldsymbol{w} \leftarrow \boldsymbol{w} + y^{(t)}\boldsymbol{x}^{(t)}$
    - $b \leftarrow b + y^{(t)}$

(Online) Perceptron Algorithm

- Initialize the parameters to all zeros:

$$\boldsymbol{\theta} = [0 \quad 0 \quad \cdots \quad 0]$$

- For $t = 1, 2, 3, \ldots$
  - Receive an unlabeled example, $\boldsymbol{x}^{(t)}$
  - Predict its label, $\hat{y} = \text{sign}\left(\boldsymbol{\theta}^T \boldsymbol{x}'^{(t)}\right) = \begin{cases} +1 \text{ if } \boldsymbol{\theta}^T \boldsymbol{x}'^{(t)} \geq 0 \\ -1 \text{ otherwise} \end{cases}$
  - Observe its true label, $y^{(t)}$
  - If we misclassified an example ($y^{(t)} \neq \hat{y}$):
    - $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + y^{(t)} \boldsymbol{x}'^{(t)}$

1 prepended to $\boldsymbol{x}^{(t)}$

Automatically handles updating the intercept

Perceptron Inductive Bias

1. Decision boundary should be linear
2. Recent mistakes are more important than older ones (and should be corrected immediately)

(Online) Perceptron Algorithm

上海科技大学 ShanghaiTech University

Algorithm 1 Perceptron Learning Algorithm (Online)

1: **procedure** PERCEPTRON($\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots\}$)
2: $\quad\boldsymbol{\theta} \leftarrow \mathbf{0}$ ▷ Initialize parameters
3: $\quad$**for** $i \in \{1, 2, \ldots\}$ **do** ▷ For each example
4: $\quad\quad\hat{y} \leftarrow \text{sign}(\boldsymbol{\theta}^T \mathbf{x}^{(i)})$ ▷ Predict
5: $\quad\quad$**if** $\hat{y} \neq y^{(i)}$ **then** ▷ If mistake
6: $\quad\quad\quad\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + y^{(i)}\mathbf{x}^{(i)}$ ▷ Update parameters
7: $\quad$**return** $\boldsymbol{\theta}$

(Batch) Perceptron Algorithm

Learning for Perceptron also works if we have a fixed training dataset, D. We call this the “batch” setting in contrast to the “online” setting that we’ve discussed so far.

Algorithm 1 Perceptron Learning Algorithm (Batch)

procedure PERCEPTRON($\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N)}, y^{(N)})\}$)
    $\boldsymbol{\theta} \leftarrow \mathbf{0}$ ▷ Initialize parameters
    while not converged **do**
        for $i \in \{1, 2, \ldots, N\}$ **do** ▷ For each example
            $\hat{y} \leftarrow \text{sign}(\boldsymbol{\theta}^T \mathbf{x}^{(i)})$ ▷ Predict
            if $\hat{y} \neq y^{(i)}$ **then** ▷ If mistake
                $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + y^{(i)} \mathbf{x}^{(i)}$ ▷ Update parameters
    return $\boldsymbol{\theta}$

(Batch) Perceptron Algorithm

Learning for Perceptron also works if we have a fixed training dataset, D. We call this the "batch" setting in contrast to the "online" setting that we've discussed so far.

Algorithm 1 Perceptron Learning Algorithm (Batch)

procedure PERCEPTRON($\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N)},$
    $\boldsymbol{\theta} \leftarrow \mathbf{0}$ ▷ Initialize para
    while not converged **do**
        for $i \in \{1, 2, \ldots, N\}$ **do** ▷ For each e
            $\hat{y} \leftarrow \text{sign}(\boldsymbol{\theta}^T \mathbf{x}^{(i)})$ ▷
            if $\hat{y} \neq y^{(i)}$ **then** ▷ If
                $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + y^{(i)} \mathbf{x}^{(i)}$ ▷ Update para
    return $\boldsymbol{\theta}$

Def: We say that the (batch) perceptron algorithm has **converged** if it stops making mistakes on the training data (perfectly classifies the training data).

Perceptron Exercise

上海科技大学
ShanghaiTech University

Question:

*Unlike Decision Trees and K-Nearest Neighbors, the Perceptron algorithm **does not suffer from overfitting** because it does not have any hyperparameters that could be over-tuned on the training data.*

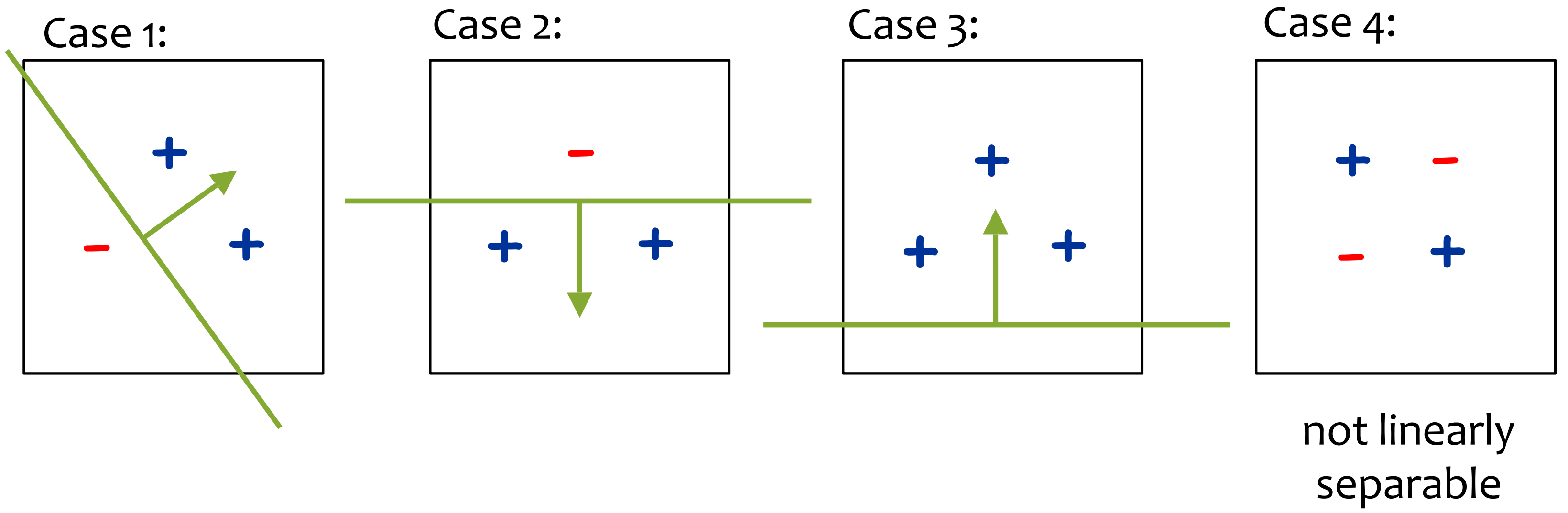
A. *True*
B. *False*
C. *True and False*

Answer:

PERCEPTRON MISTAKE BOUND

Definitions

Def: For a **binary classification** problem, a set of examples S is **linearly separable** if there exists a linear decision boundary that can separate the points



Definitions

上海科技大学 ShanghaiTech University

Def: The **margin** γ for a dataset D is the greatest possible distance between a linear separator and the closest data point in D to that linear separator

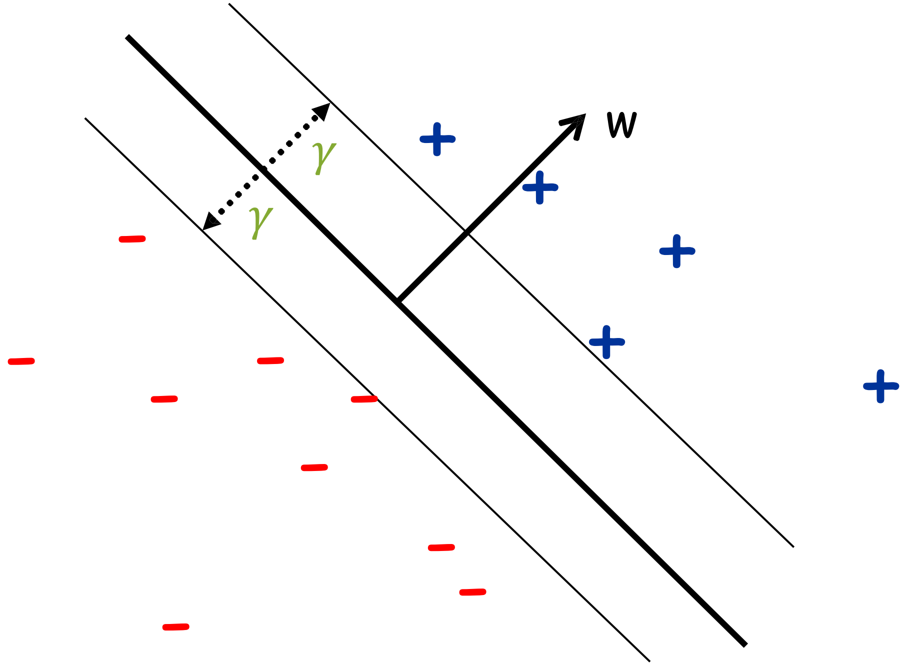


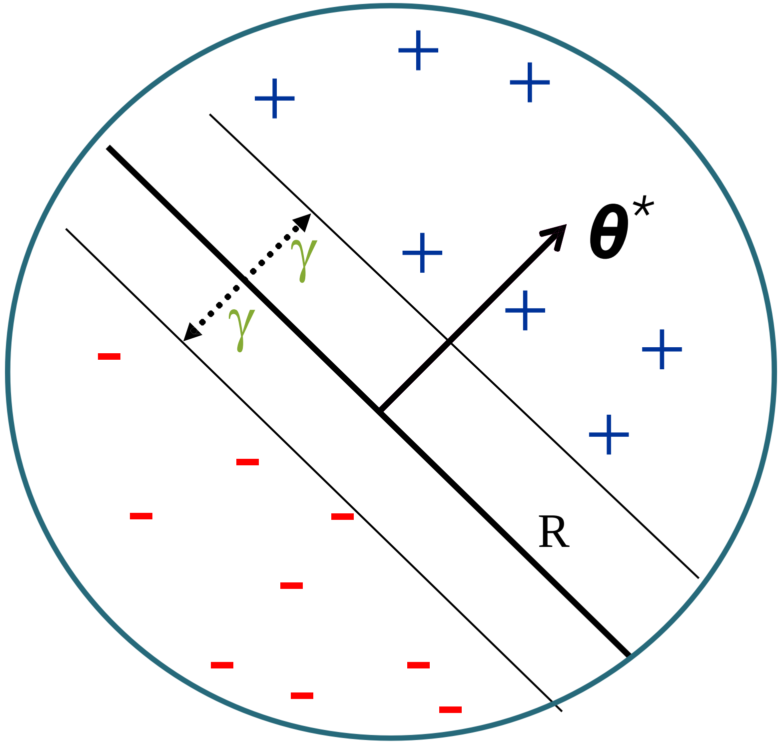
Figure from Nina Balcan

Perceptron Mistake Bound

上海科技大学
ShanghaiTech University

Guarantee: if some data has margin γ and all points lie inside a ball of radius R rooted at the origin, then the online Perceptron algorithm makes $\leq (R/\gamma)^2$ mistakes

(Normalized margin: multiplying all points by 100, or dividing all points by 100, doesn't change the number of mistakes! The algorithm is invariant to scaling.)

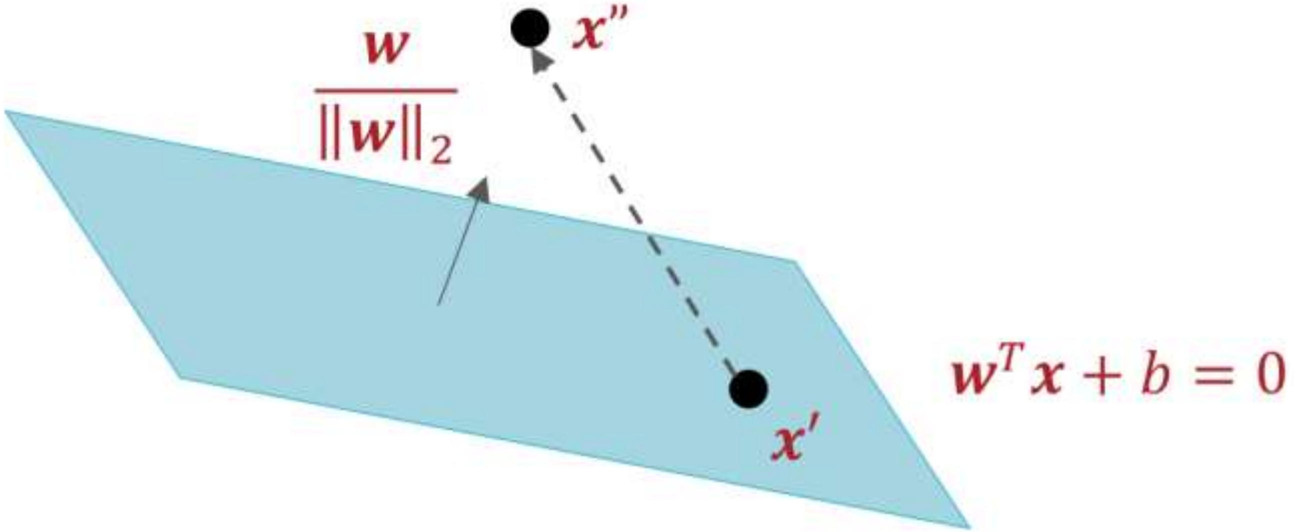


Main Takeaway: For **linearly separable** data, if the perceptron algorithm cycles repeatedly through the data, it will **converge** in a finite # of steps.

Slide adapted from Nina Balcan

Computing the Margin

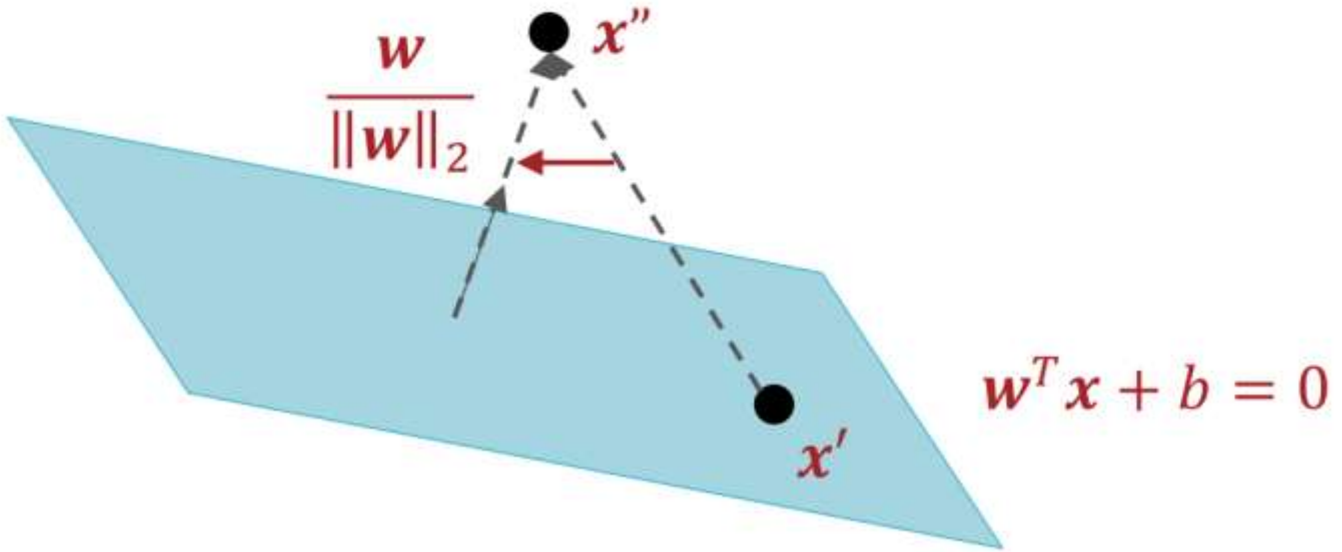
- Let $\boldsymbol{x}'$ be an arbitrary point on the hyperplane $\mathbf{w}^T\boldsymbol{x} + b = 0$ and let $\boldsymbol{x}''$ be an arbitrary point
- The distance between $\boldsymbol{x}''$ and $\mathbf{w}^T\boldsymbol{x} + b = 0$ is equal to the magnitude of the projection of $\boldsymbol{x}'' - \boldsymbol{x}'$ onto $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the unit vector orthogonal to the hyperplane



Computing the Margin

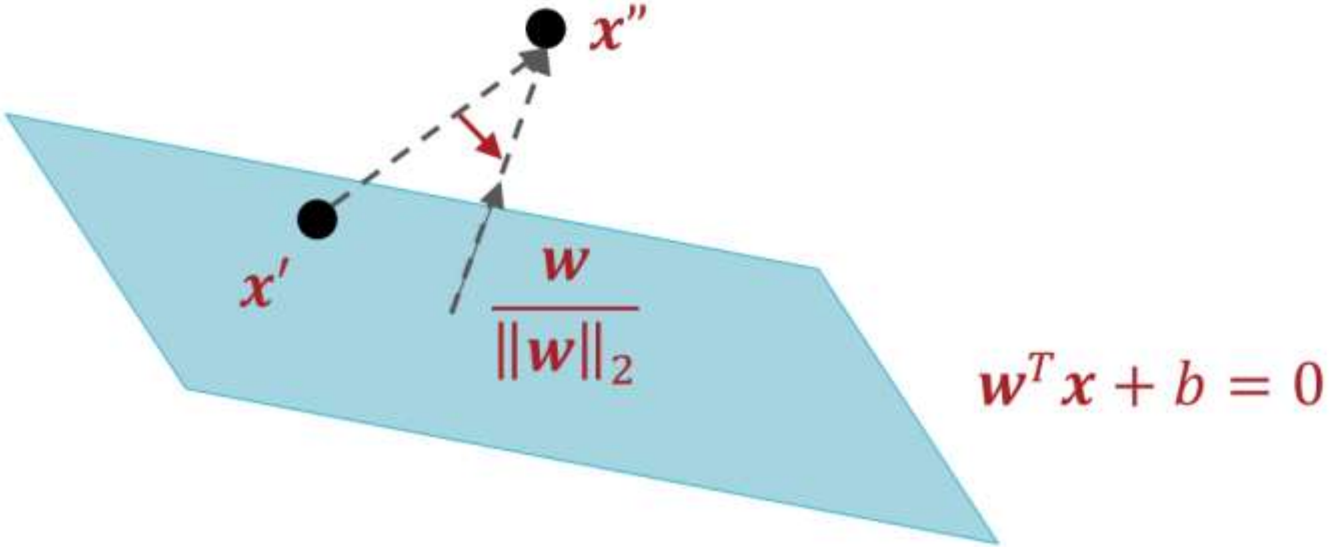
上海科技大学 ShanghaiTech University

- Let $\boldsymbol{x}'$ be an arbitrary point on the hyperplane $\mathbf{w}^T\boldsymbol{x} + b = 0$ and let $\boldsymbol{x}''$ be an arbitrary point
- The distance between $\boldsymbol{x}''$ and $\mathbf{w}^T\boldsymbol{x} + b = 0$ is equal to the magnitude of the projection of $\boldsymbol{x}'' - \boldsymbol{x}'$ onto $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the unit vector orthogonal to the hyperplane



Computing the Margin

- Let $\boldsymbol{x}'$ be an arbitrary point on the hyperplane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ and let $\boldsymbol{x}''$ be an arbitrary point
- The distance between $\boldsymbol{x}''$ and $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ is equal to the magnitude of the projection of $\boldsymbol{x}'' - \boldsymbol{x}'$ onto $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the unit vector orthogonal to the hyperplane



Computing the Margin

上海科技大学 ShanghaiTech University

- Let $\boldsymbol{x}'$ be an arbitrary point on the hyperplane and let $\boldsymbol{x}''$ be an arbitrary point
- The distance between $\boldsymbol{x}''$ and $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ is equal to the magnitude of the projection of $\boldsymbol{x}'' - \boldsymbol{x}'$ onto $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, the unit vector orthogonal to the hyperplane

$$\left|\frac{\boldsymbol{w}^T(\boldsymbol{x}'' - \boldsymbol{x}')}{\|\boldsymbol{w}\|_2}\right| = \frac{|\boldsymbol{w}^T\boldsymbol{x}'' - \boldsymbol{w}^T\boldsymbol{x}'|}{\|\boldsymbol{w}\|_2} = \frac{|\boldsymbol{w}^T\boldsymbol{x}'' + b|}{\|\boldsymbol{w}\|_2}$$

PROOF OF THE MISTAKE BOUND

Analysis: Perceptron

上海科技大学 ShanghaiTech University

Perceptron Mistake Bound

Theorem 0.1 (Block (1962), Novikoff (1962)).
Given dataset: $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$.
Suppose:

1. *Finite size inputs:* $||x^{(i)}|| \leq R$
2. *Linearly separable data:* $\exists \boldsymbol{\theta}^*$ s.t. $||\boldsymbol{\theta}^*|| = 1$ *and* $y^{(i)}(\boldsymbol{\theta}^* \cdot \mathbf{x}^{(i)}) \geq \gamma, \forall i$ *and some* $\gamma > 0$

Then: The number of mistakes made by the Perceptron algorithm on this dataset is

$$k \leq (R/\gamma)^2$$

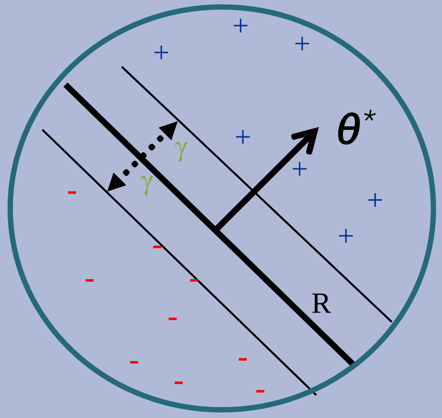


Figure from Nina Balcan

Analysis: Percept

Common Misunderstanding: The **radius** is **centered at the origin**, not at the center of the points.

Perceptron Mistake Boun

Theorem 0.1 (Block (1962), Novikoff (19

Given dataset: $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$

Suppose:

1. *Finite size inputs:* $||x^{(i)}|| \leq R$
2. *Linearly separable data:* $\exists \boldsymbol{\theta}^*$ *s.t.* $||\boldsymbol{\theta}^*|| = 1$ *and* $y^{(i)}(\boldsymbol{\theta}^* \cdot \mathbf{x}^{(i)}) \geq \gamma, \forall i$ *and some* $\gamma > 0$

Then: The number of mistakes made by the Perceptron algorithm on this dataset is

$$k \leq (R/\gamma)^2$$

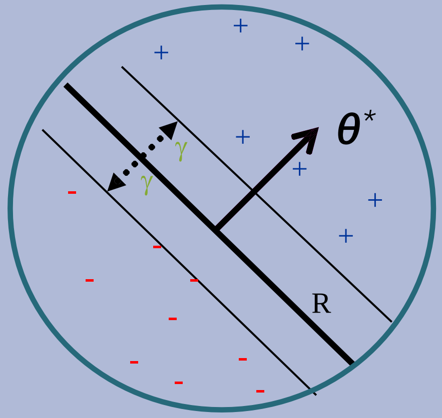
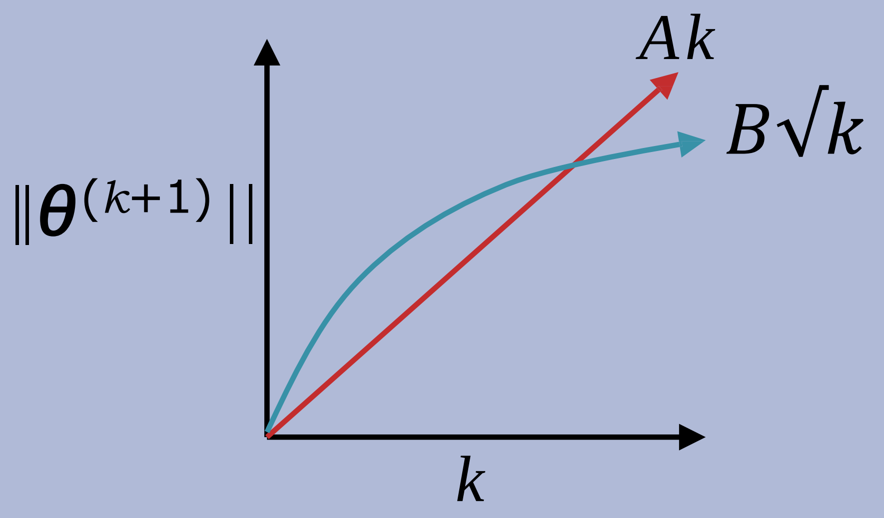


Figure from Nina Balcan

Proof of Perceptron Mistake Bound:

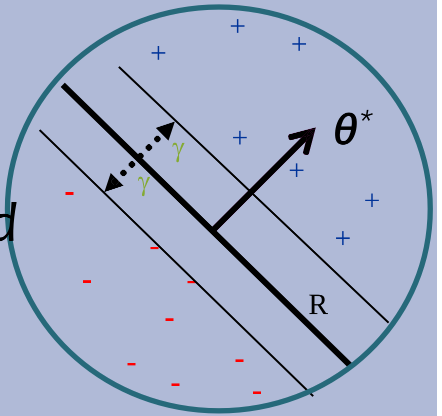
We will show that there exist constants A and B s.t.

$$Ak \leq \|\boldsymbol{\theta}^{(k+1)}\| \leq B\sqrt{k}$$



Theorem 0.1 (Block (1962), Novikoff (1962)).
Given dataset: $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$.
Suppose:

1. *Finite size inputs:* $||x^{(i)}|| \leq R$
2. *Linearly separable data:* $\exists\ \boldsymbol{\theta}^*$ *s.t.* $||\boldsymbol{\theta}^*|| = 1$ *and* $y^{(i)}(\boldsymbol{\theta}^* \cdot \mathbf{x}^{(i)}) \geq \gamma, \forall i$ *and some* $\gamma > 0$



Then: The number of mistakes made by the Perceptron algorithm on this dataset is

$$k \leq (R/\gamma)^2$$

Algorithm 1 Perceptron Learning Algorithm (Online)

procedure PERCEPTROĒ($\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \ldots\}$)
  $\boldsymbol{\theta} \leftarrow \mathbf{0}$, $k = 1$ — D Initialize parameters
  for $i \in \{1, 2, \ldots\}$ **do** — D For each example
    if $y^{(i)}(\boldsymbol{\theta}^{(k)} \cdot \mathbf{x}^{(i)}) \leq 0$ **then** — D If mistake
      $\boldsymbol{\theta}^{(k+1)} \leftarrow \boldsymbol{\theta}^{(k)} + y^{(i)}\mathbf{x}^{(i)}$ — D Update parameters
      $k \leftarrow k + 1$
  return $\boldsymbol{\theta}$

Proof of Perceptron Mistake Bound:

Part 1: for some A, $Ak \leq \|\boldsymbol{\theta}^{(k+1)}\|$

$$\boldsymbol{\theta}^{(k+1)} \cdot \boldsymbol{\theta}^{*} = (\boldsymbol{\theta}^{(k)} + y^{(i)}\mathbf{x}^{(i)})\boldsymbol{\theta}^{*}$$

by Perceptron algorithm update

$$= \boldsymbol{\theta}^{(k)} \cdot \boldsymbol{\theta}^{*} + y^{(i)}(\boldsymbol{\theta}^{*} \cdot \mathbf{x}^{(i)})$$

$$\geq \boldsymbol{\theta}^{(k)} \cdot \boldsymbol{\theta}^{*} + \gamma$$

by assumption

$$\Rightarrow \boldsymbol{\theta}^{(k+1)} \cdot \boldsymbol{\theta}^{*} \geq k\,\gamma$$

by induction on k since $\theta^{(1)} = \mathbf{0}$

$$\Rightarrow \|\boldsymbol{\theta}^{(k+1)}\| \geq k\,\gamma$$

since $\underbrace{||\mathbf{r}|| \times ||\mathbf{m}|| \geq \mathbf{r} \cdot \mathbf{m}}_{\text{Cauchy-Schwartz inequality}}$ and $||\theta^{*}|| = 1$

Proof of Perceptron Mistake Bound

Part 2: for some B, $\|\boldsymbol{\theta}^{(k+1)}\| \leq B\sqrt{k}$

$\|\boldsymbol{\theta}_{(k+1)}\|_2 = \|\boldsymbol{\theta}_{(k)} + y_{(i)}\mathbf{x}_{(i)}\|_2$

by Perceptron algorithm update

$= \|\boldsymbol{\theta}^{(k)}\|^2 + (y^{(i)})^2\|\mathbf{x}^{(i)}\|^2 + 2y^{(i)}(\boldsymbol{\theta}^{(k)} \cdot \mathbf{x}^{(i)})$

$\leq \|\boldsymbol{\theta}^{(k)}\|^2 + (y^{(i)})^2\|\mathbf{x}^{(i)}\|^2$

since kth mistake $\Rightarrow y^{(i)}(\boldsymbol{\theta}^{(k)} \cdot \mathbf{x}^{(i)}) \leq 0$

$= \|\boldsymbol{\theta}^{(k)}\|^2 + R^2$

since $(y^{(i)})^2\|\mathbf{x}^{(i)}\|^2 = \|\mathbf{x}^{(i)}\|^2 = R^2$ by assumption and $(y^{(i)})^2 = 1$

$\Rightarrow \|\boldsymbol{\theta}^{(k+1)}\|^2 \leq kR^2$

by induction on k since $(\theta^{(1)})^2 = 0$

$\Rightarrow \|\boldsymbol{\theta}^{(k+1)}\| \leq \sqrt{k}\, R$

Proof of Perceptron Mistake Bound:

Part 3: Combining the bounds finishes the proof.

$$k\gamma \leq ||\boldsymbol{\theta}^{(k+1)}|| \leq \sqrt{k}R$$

$$\Rightarrow k \leq \underbrace{(R/\gamma)^2}$$

The total number of mistakes must be less than this

What if the data is *not* linearly separable?

1. Perceptron will **not converge** in this case (it can't!)
2. However, Freund & Schapire (1999) show that by projecting the points (hypothetically) into a higher dimensional space, we can achieve a similar bound on the number of mistakes made on **one pass** through the sequence of examples

Theorem 2. *Let* $\langle(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\rangle$ *be a sequence of labeled examples with* $\|\mathbf{x}_i\| \leq R$. *Let* $\mathbf{u}$ *be any vector with* $\|\mathbf{u}\| = 1$ *and let* $\gamma > 0$. *Define the deviation of each example as*

$$d_i = \max\{0, \gamma - y_i(\mathbf{u} \cdot \mathbf{x}_i)\},$$

and define $D = \sqrt{\sum_{i=1}^m d_i^2}$. *Then the number of mistakes of the online perceptron algorithm on this sequence is bounded by*

$$\left(\frac{R + D}{\gamma}\right)^2.$$

Summary: Perceptron

上海科技大学 ShanghaiTech University

- Perceptron is a **linear classifier**
- **Simple learning algorithm**: when a mistake is made, add / subtract the features
- Perceptron will converge if the data are **linearly separable**, it will **not** converge if the data are **linearly inseparable**
- For linearly separable and inseparable data, we can **bound the number of mistakes** (geometric argument)
- **Extensions** support nonlinear separators and structured prediction

Perceptron Learning Objectives

You should be able to…

- Explain the difference between online learning and batch learning
- Implement the perceptron algorithm for binary classification [CIML]
- Determine whether the perceptron algorithm will converge based on properties of the dataset, and the limitations of the convergence guarantees
- Describe the inductive bias of perceptron and the limitations of linear models
- Draw the decision boundary of a linear model
- Identify whether a dataset is linearly separable or not
- Defend the use of a bias term in perceptron

Extensions of Perceptron

- Voted Perceptron
  - generalizes better than (standard) perceptron
  - memory intensive (keeps around every weight vector seen during training, so each one can vote)
- Averaged Perceptron
  - empirically similar performance to voted perceptron
  - can be implemented in a memory efficient way (running averages are efficient)
- Kernel Perceptron
  - Choose a kernel K(x', x)
  - Apply the **kernel trick** to Perceptron
  - Resulting algorithm is **still very simple**
- Structured Perceptron
  - Basic idea can also be applied when **y** ranges over an exponentially large set
  - Mistake bound **does not** depend on the size of that set