Lecture 9: Linear Models

COMP 411, Fall 2021
Victoria Manfredi

WESLEYAN UNIVERSITY

Acknowledgements: These slides are based primarily on those created by Vivek Srikumar (Utah) and Dan Roth (Penn), and lectures by Balaraman Ravindran (IIT Madras)

Today's Topics

Homework 4 out

- Due Thursday, October 7 by 11:59p

Linear models

- Overview
- Geometry of linear classifiers
- A notational simplification
- Learning linear classifiers
- Expressivity

Linear models

OVERVIEW

Checkpoint: the bigger picture

Supervised learning: instances, concepts, and hypotheses

- Labeled data $\rightarrow$ Learning algorithm$\rightarrow$ Hypothesis/Model h
- New example $\rightarrow h \rightarrow$Prediction

Specific learners

- Decision trees

General ML ideas

- Features as high dimensional vectors
- Overfitting

Is learning possible at all?

There are $2^{16} = 65536$ possible Boolean functions over 4 inputs

- Why? There are 16 possible outputs. each way to fill these 16 slots is a different function, giving 2^{16} functions

We have seen 7 outputs

We *cannot* know what the rest are without seeing them

- Think of an adversary filing in the labels every time you make a guess at a function

x_1	x_2	x_3	x_4	y
0	0	0	0	?
0	0	0	1	?
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	?
1	0	0	0	?
1	0	0	1	1
1	0	1	0	?
1	0	1	1	?
1	1	0	0	0
1	1	0	1	?
1	1	1	0	?
1	1	1	1	?

Is learning possible at all?

There are $2^{16} = 65536$ possible Boolean functions over 4 inputs

- Why? There are 16 possible outputs. each way to fill these 16 slots is a d
fu

We hav

How could we possibly learn anything?

We *cannot* know what the rest are without seeing them

- Think of an adversary filing in the labels every time you make a guess at a function

x_1	x_2	x_3	x_4	y
0	0	0	0	?
0	0	0	1	?
0	0	1	0	0
1	0	0	0	?
1	0	0	1	1
1	0	1	0	?
1	0	1	1	?
1	1	0	0	0
1	1	0	1	?
1	1	1	0	?
1	1	1	1	?

Solution: restrict the search space

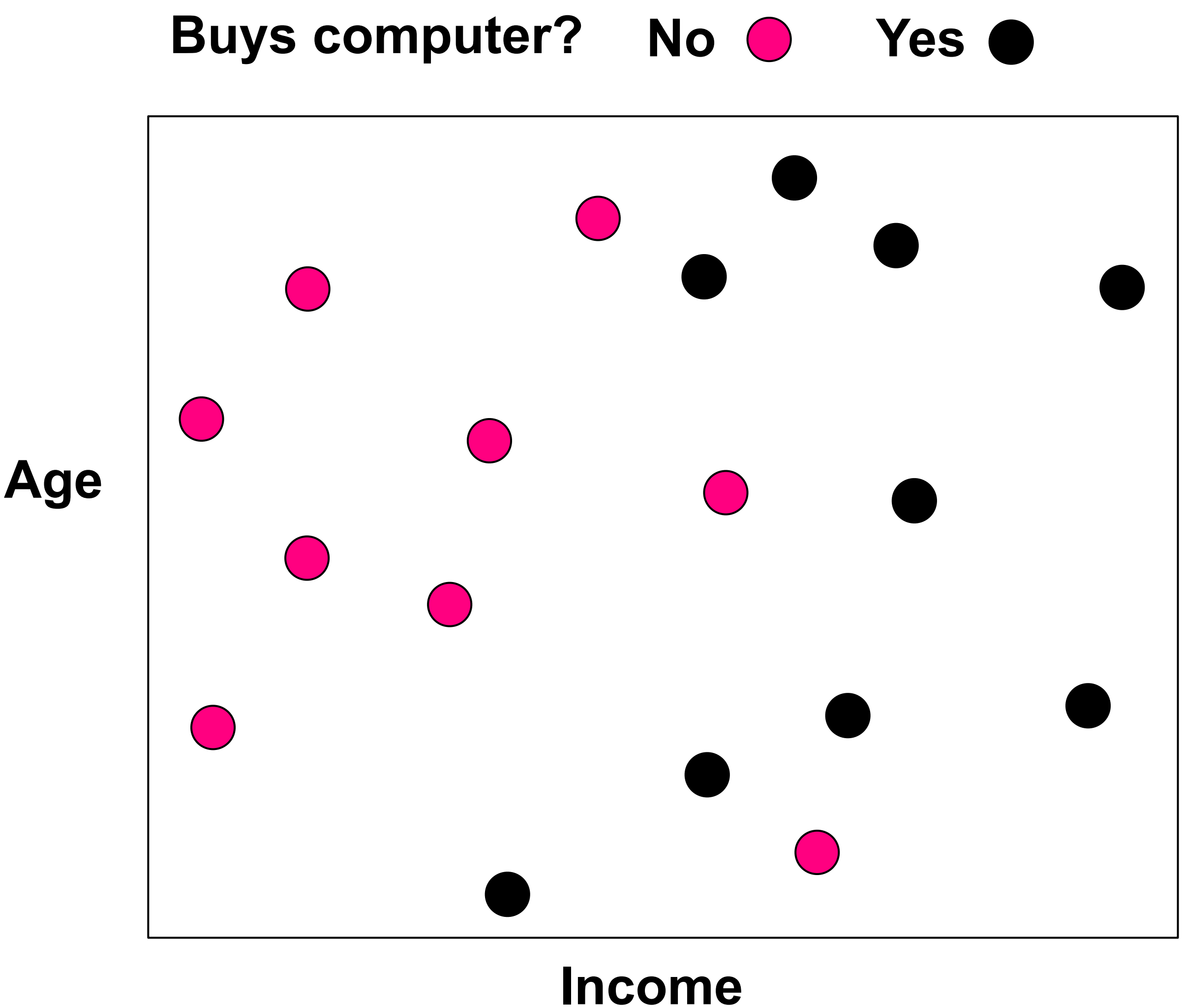
A hypothesis space is the set of possible functions we consider

We were looking at the space of all Boolean functions. Instead we choose a hypothesis space that is smaller than the space of all functions

For example:

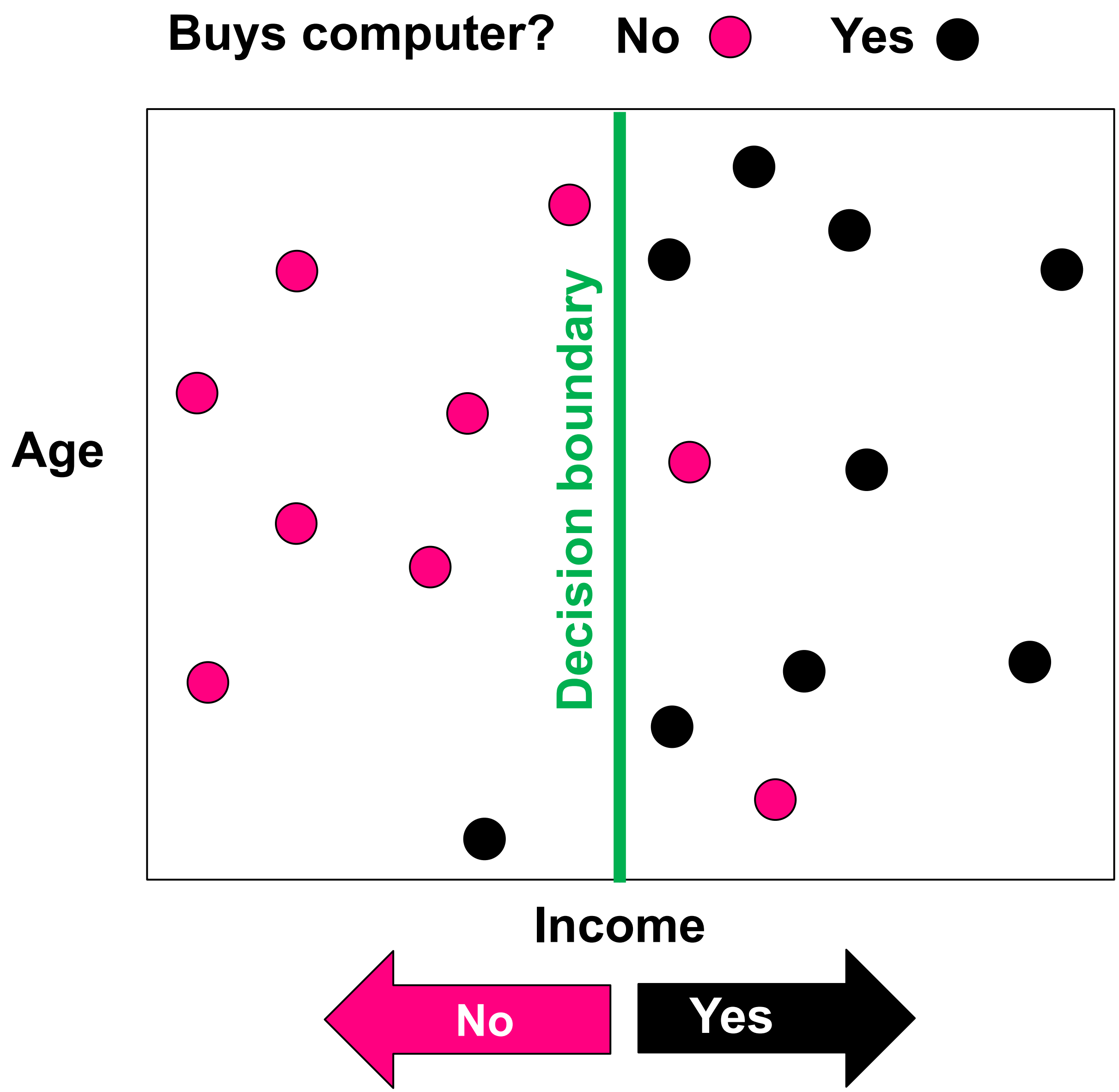
- Only simple conjunctions with 4 variables, there are 16 conjunctions without negations
- Simple disjunction
- m-of-n rules: Fix a set of n variables. At least m of them must be true
- Linear functions
- …

Training set for classification



Q: How should we draw decision boundary separating No and Yes?
I.e., which is the ***better classifier?***

Classifier attempt 1



Function

- if person's income ≤ x, then person will not buy a computer.
- if person's income > x then person will buy a computer

Problem

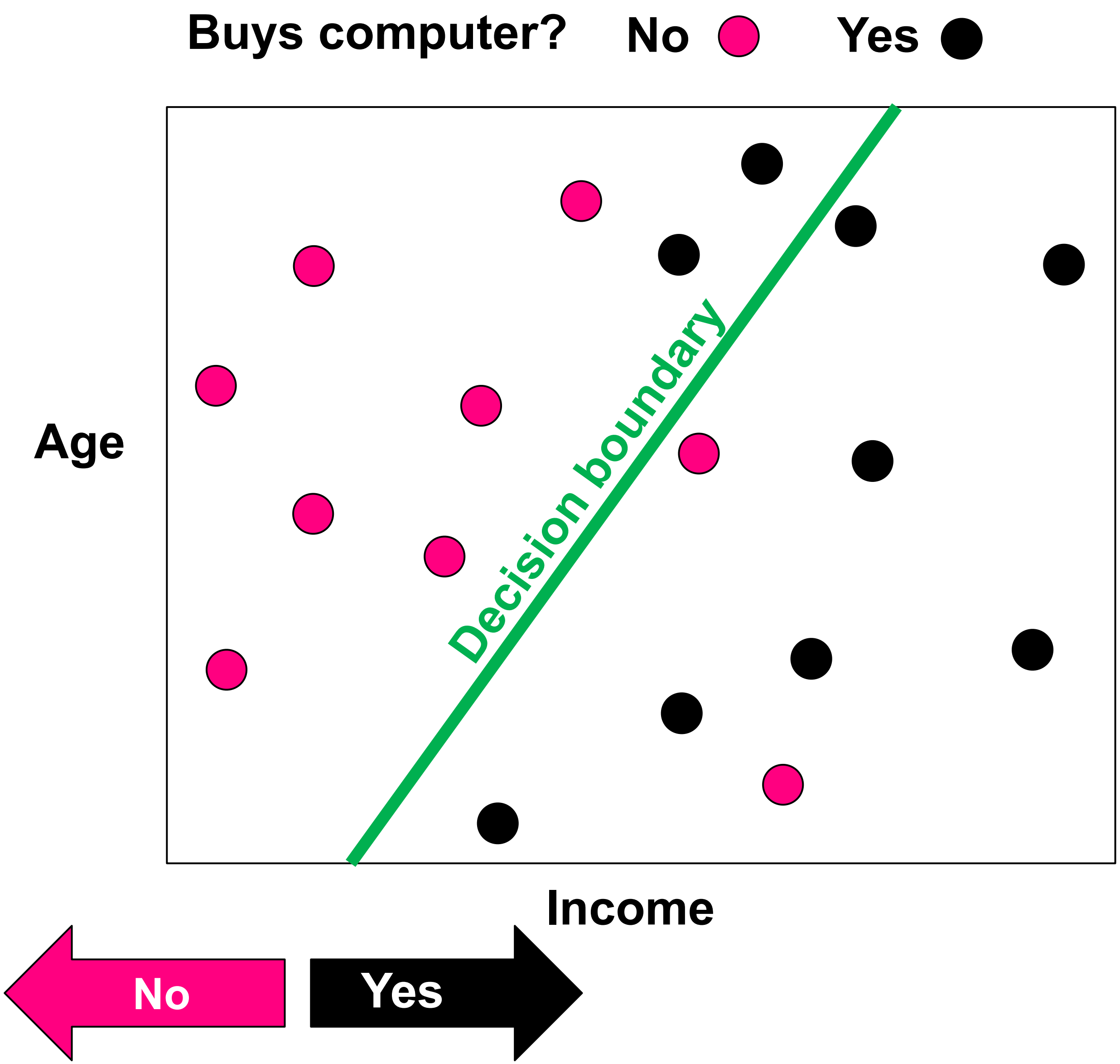
- We ignored age …

Question

- Can we do better?

vumanfredi@wesleyan.edu

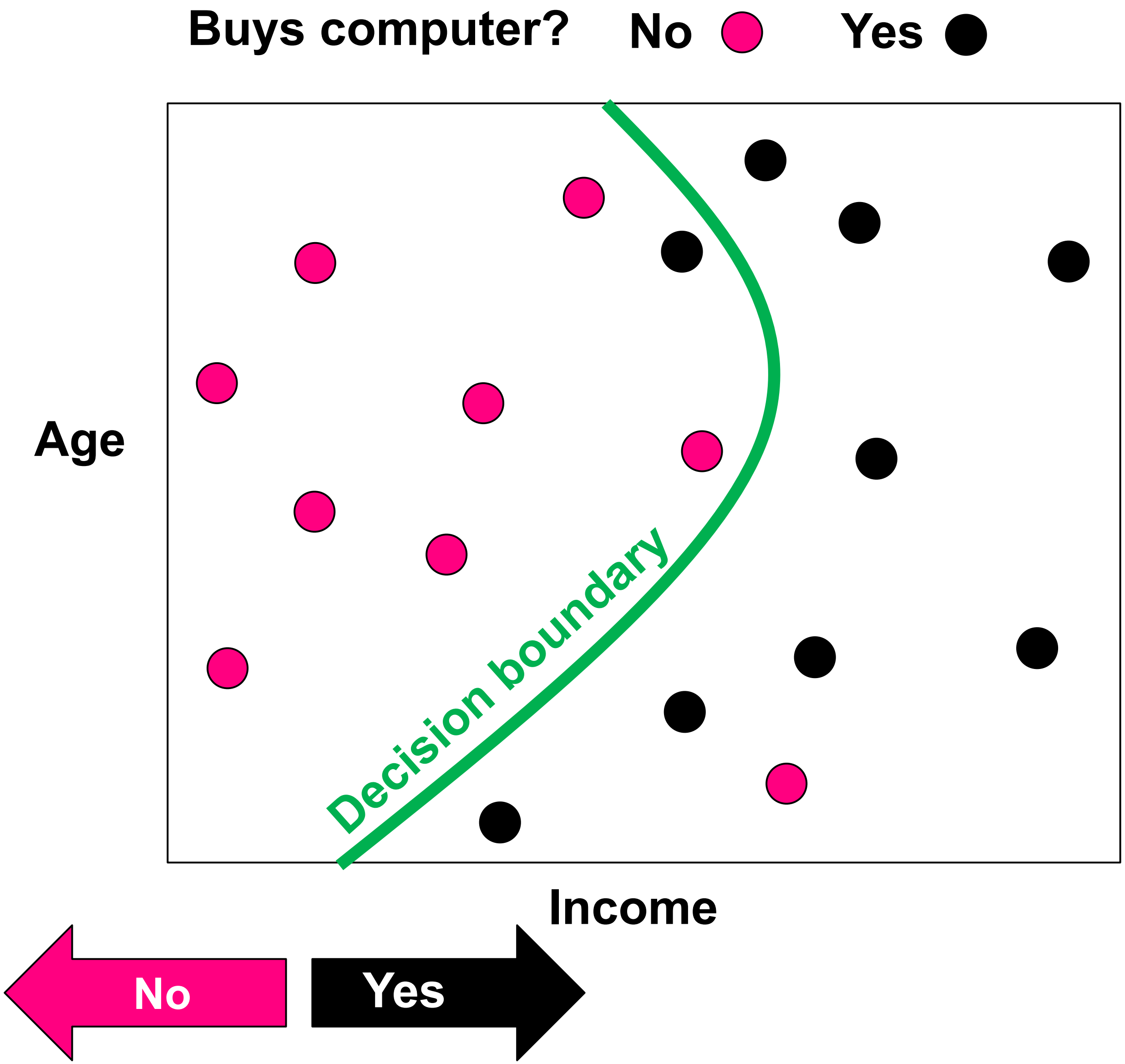
Classifier attempt 2



Yes: improve performance at the cost of paying attention to age

Q: Can we do even better?

Classifier attempt 3



Yes: almost everything correct at cost of more complex classifier

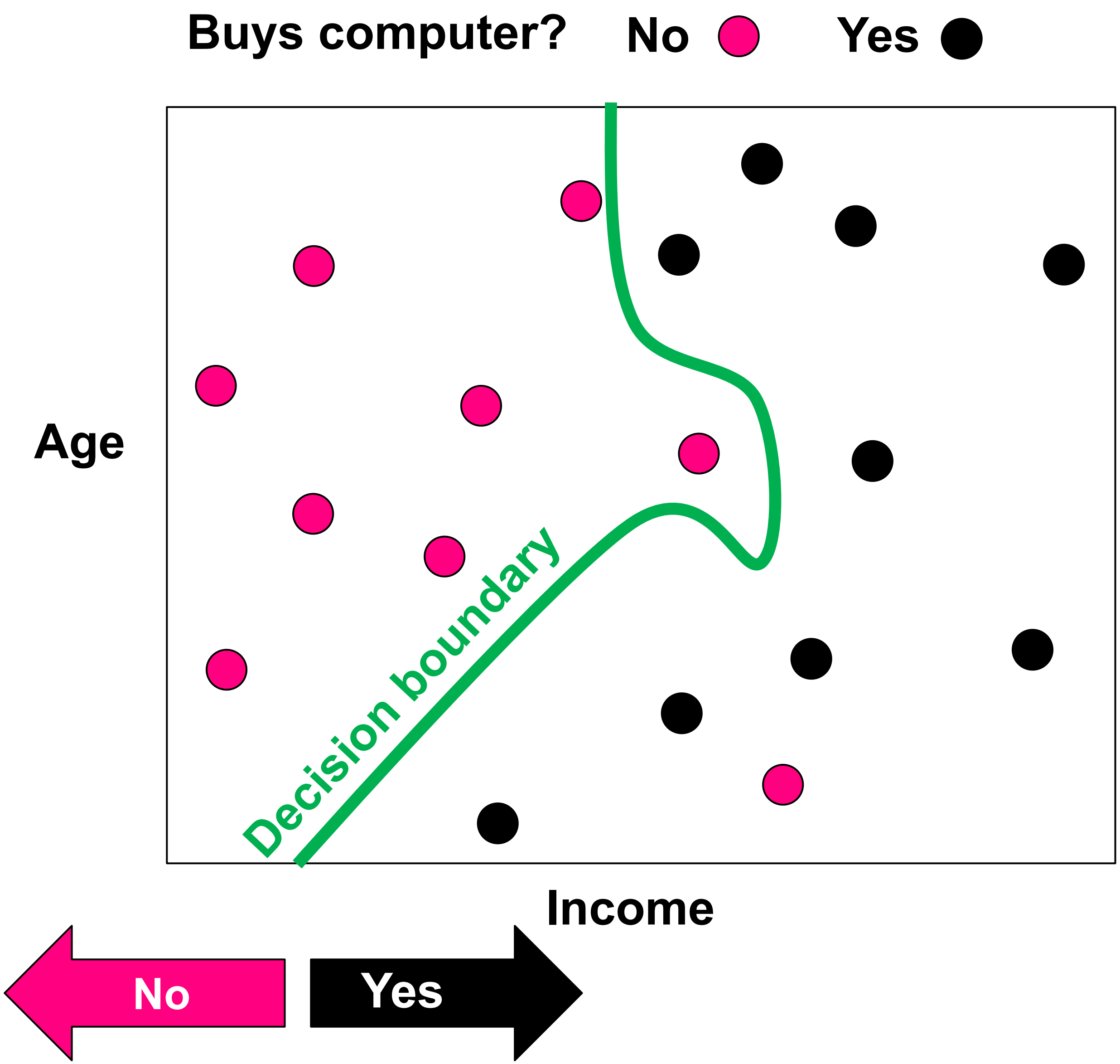
So far:

1. $y = b$
2. $y = mx + b$
3. $y = ax^2 + bx + c$

Q: Can we do even better?

vumanfredi@wesleyan.edu

Classifier attempt 4

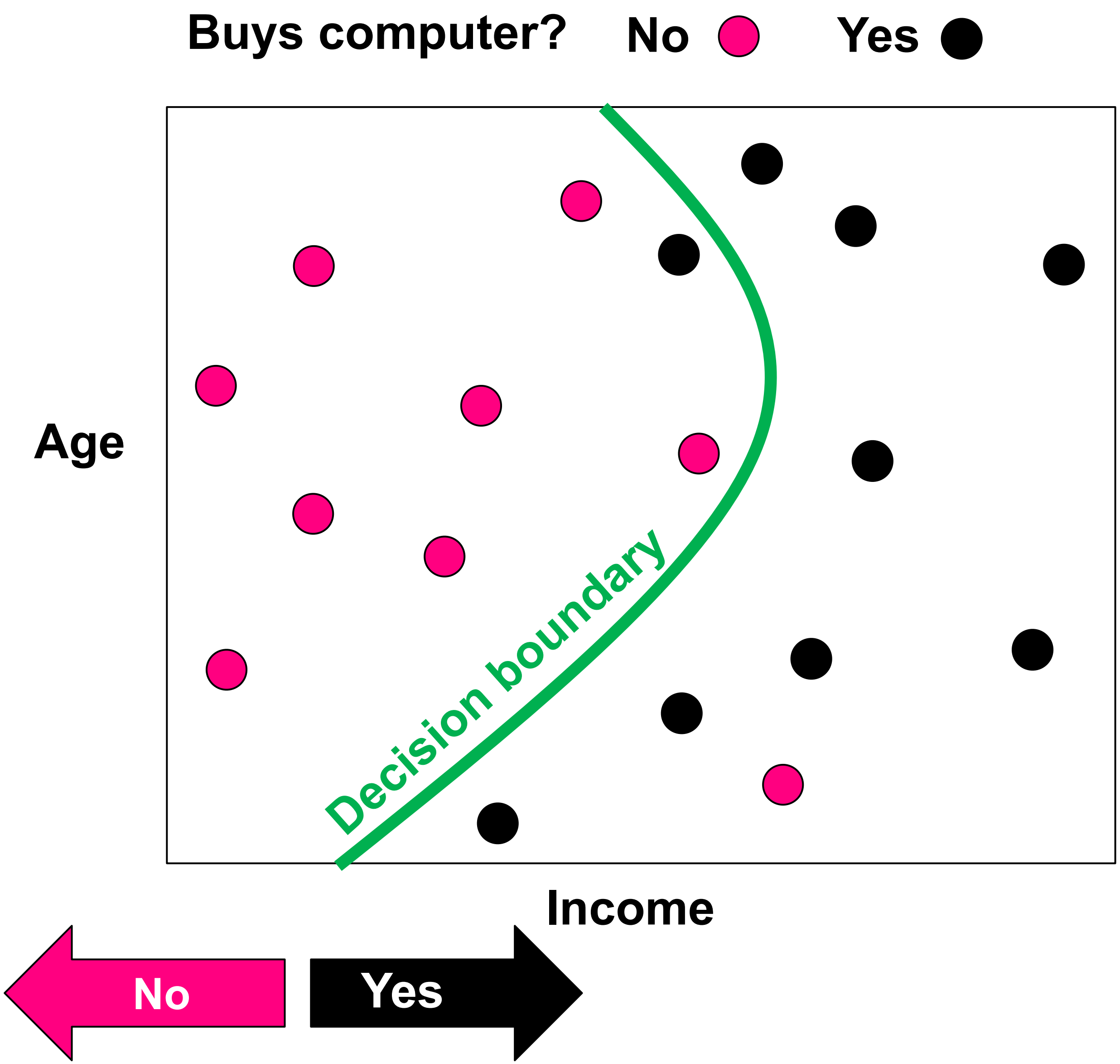


Uh-oh: this doesn't seem right! ***<u>Risks overfitting</u>***

Errant pink point

- Data mis-recorded?
- Person got emergency call and left shop?
- Other noise …

The winner!



Why? Trades-off complexity vs. accuracy

Other considerations: is there noise in the data? If so, how do we handle the noise

Linear classification vs. regression

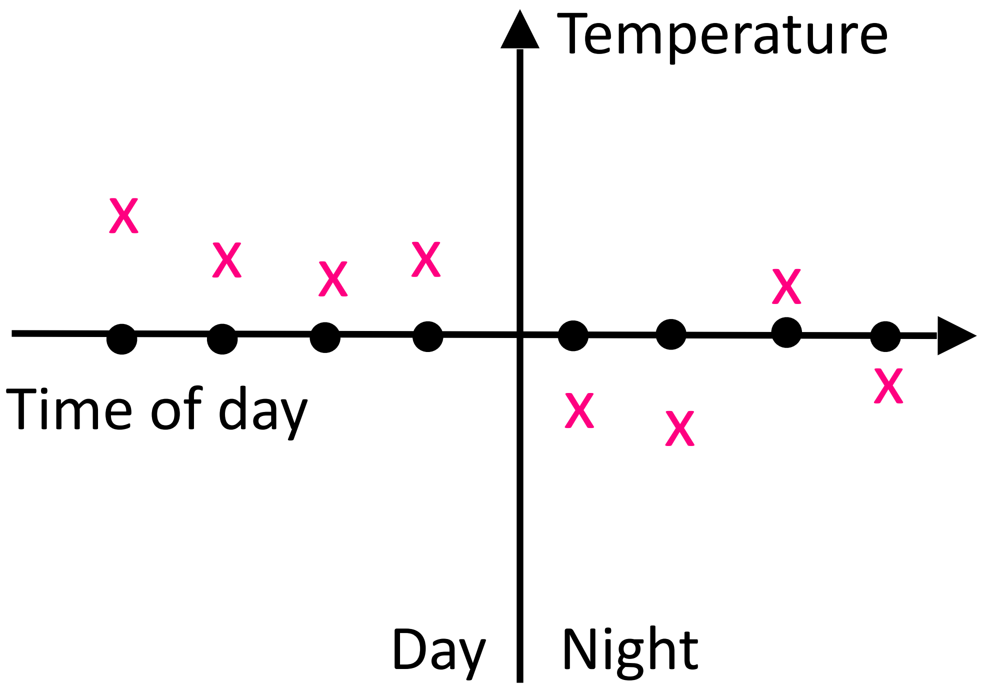
Linear classification is about predicting a discrete class label

- +1 or -1
- SPAM or NOT-SPAM
- Or more than two categories

Linear regression is about predicting real valued outputs

Training set for regression

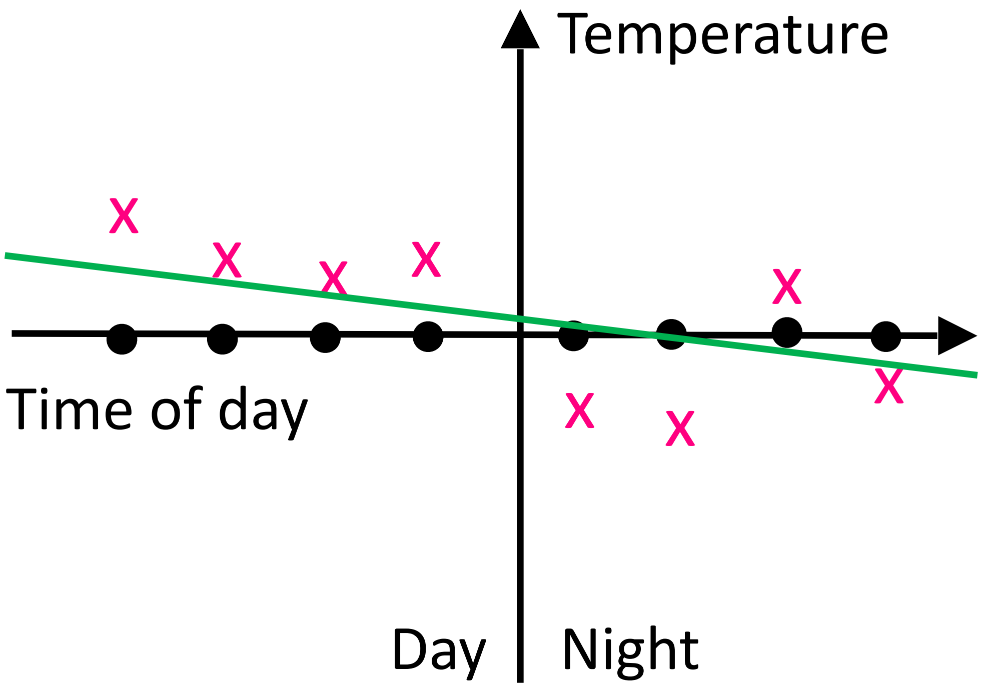
Input: time of day
Output: temperature at time



Output is no longer a discrete value: now continuous

Fit attempt 1

Input: time of day
Output: temperature at time



Function

- if person’s income ≤ x, then person will not buy a computer.
- if person’s income > x then person will buy a computer

Problem

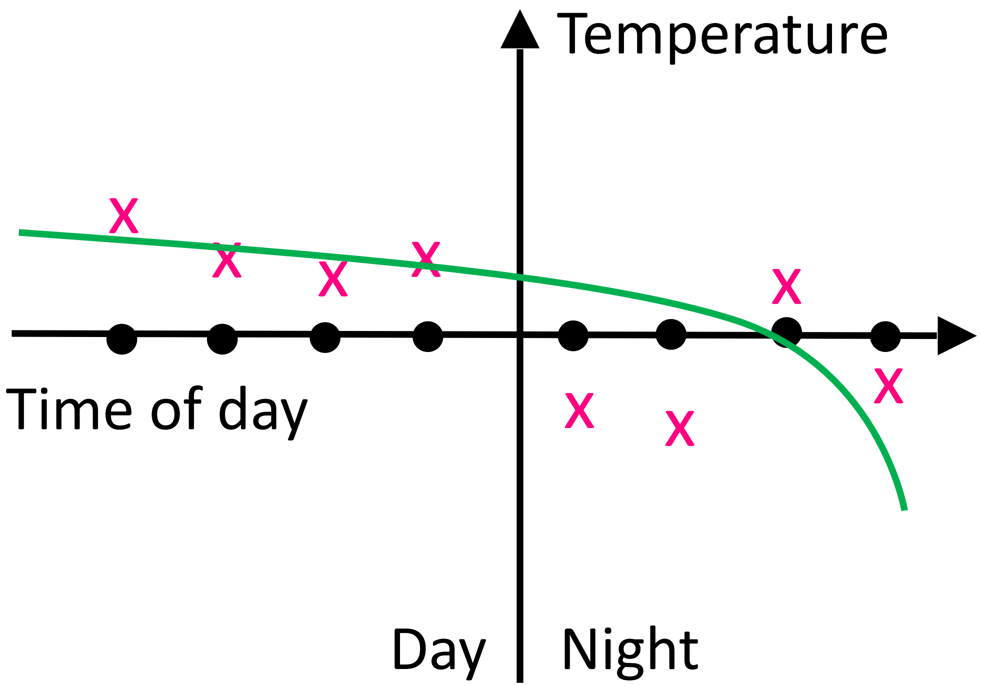
- We ignored age …

Question

- Can we do better?

Fit attempt 2

Input: time of day
Output: temperature at time



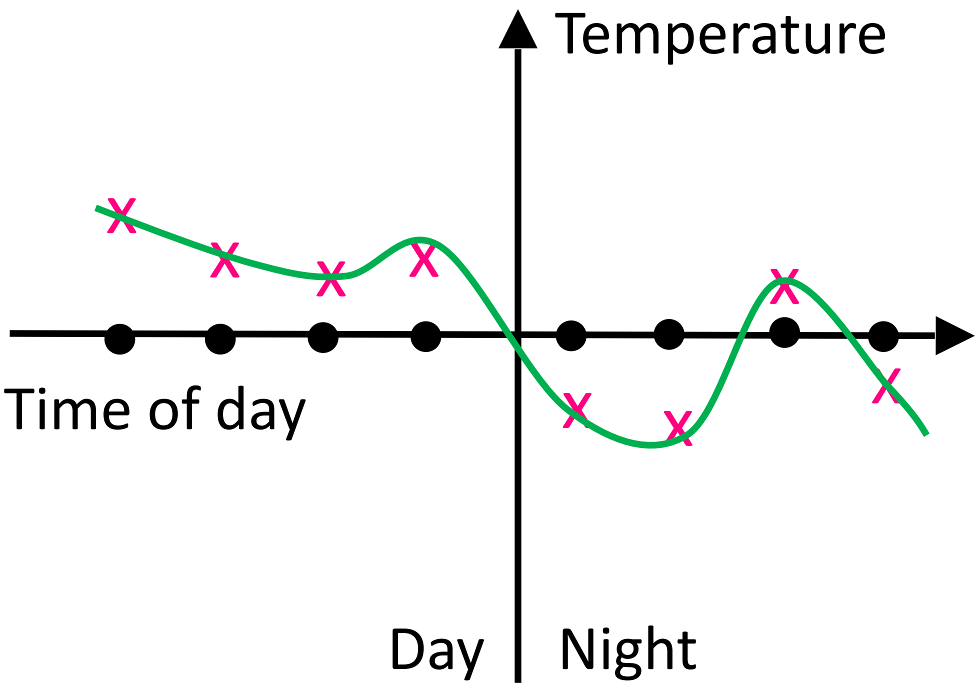
Function
- if person's income ≤ x, then person will not buy a computer.
- if person's income > x then person will buy a computer

Problem
- We ignored age …

Question
- Can we do better?

Fit attempt 3

Input: time of day
Output: temperature at time



Function
- if person’s income ≤ x, then person will not buy a computer.
- if person’s income > x then person will buy a computer

Problem
- Overfitting

Linear Classifiers

OVERVIEW

Linear classifiers: an example

Suppose we want to determine whether a robot arm is defective or not using two measurements:

1. The maximum distance the arm can reach d
2. The maximum angle it can rotate a

Suppose we use a linear decision rule that predicts defective if
$2d + 0.01a \geq 7$

We can apply this rule if we have the two measurements

For example: for a certain arm, if $d = 3$ and $a = 200$ then

$$2d + 0.01a = 8 \geq 7$$

The arm would b labeled as not defective

Linear classifiers: an example

Suppose we want to determine whether a robot arm is defective or not using two measurements:

1. The maximum distance the arm can reach d
2. The maximum angle it can rotate a

Suppose we use a linear decision rule that predicts defective if

$$2d + 0.01a \geq 7$$

This rule is an example of a linear classifier

Features are weighted and added up, the sum is checked against a threshold

Linear classifiers

Inputs are d dimensional vectors, denoted by $\mathbf{x}$

Output is a label $y \in \{-1, 1\}$

Linear Threshold Units classify an example $\mathbf{x}$ using parameters $\mathbf{w}$ (a d dimensional vector) and b (a real number) according to the following classification rule

Output = sign($\mathbf{w}^T\mathbf{x} + b$) = sign($\sum_i w_i x_i + b$)

if $\mathbf{w}^T\mathbf{x} + b \geq 0 \Rightarrow y = +1$

if $\mathbf{w}^T\mathbf{x} + b < 0 \Rightarrow y = -1$

b is called the ***bias*** term

Standard form of a line

$$Ax + By = C$$

A, B, and C are real numbers

A and B are not both zero

Drawing line

$\mathbf{w} \cdot \mathbf{x} + b$

2-dimensions: $w_1x_1 + w_2x_2 + b = 0$

Solve for x_1-intercept: $x_1 = \dfrac{-(b - w_2x_2)}{w_1}$ if $y = 0$ then $x_1 = \dfrac{-b}{w_1}$

Solve for x_2-intercept: $x_2 = \dfrac{-(b - w_1x_1)}{w_2}$if $y = 0$ then $x_2 = \dfrac{-b}{w_2}$

Two points: $(0, -b/w_2)$, $(-b/w_1, 0)$

Slope $= \dfrac{-b/w_2}{b/w_1}$, intercept $x_2 = \dfrac{-b}{w_2}$

Dot product

The dot product of two vectors is written as $\mathbf{m}^T\mathbf{x}$ or $\mathbf{m}\cdot\mathbf{x}$, which is defined as: $\mathbf{m}^T\mathbf{x} = \sum_{i=1}^{k} m_i x_i$

Example

$$\mathbf{m} = < 5.13, 1.08, -0.03, 7.29 >$$

$$\mathbf{x} = < x_1, x_2, x_3, x_4 >$$

$$\mathbf{m}^T\mathbf{x} = 5.13x_1 + 1.08x_2 - 0.03x_3 + 7.29x_4$$

If dot product of two vectors is zero: means the two vectors are perpendicular (90° angle)

Length or norm of a vector

The length or norm of a vector $\mathbf{v}$ is the square root of length $= ||\mathbf{v}|| = \sqrt{\mathbf{v} \cdot \mathbf{v}}$

Dot product here is not zero ($\mathbf{v}$ is not perpendicular to itself), so now have $0°$ angle: dot product of $\mathbf{v} \cdot \mathbf{v}$ gives length of $\mathbf{v}$ squared

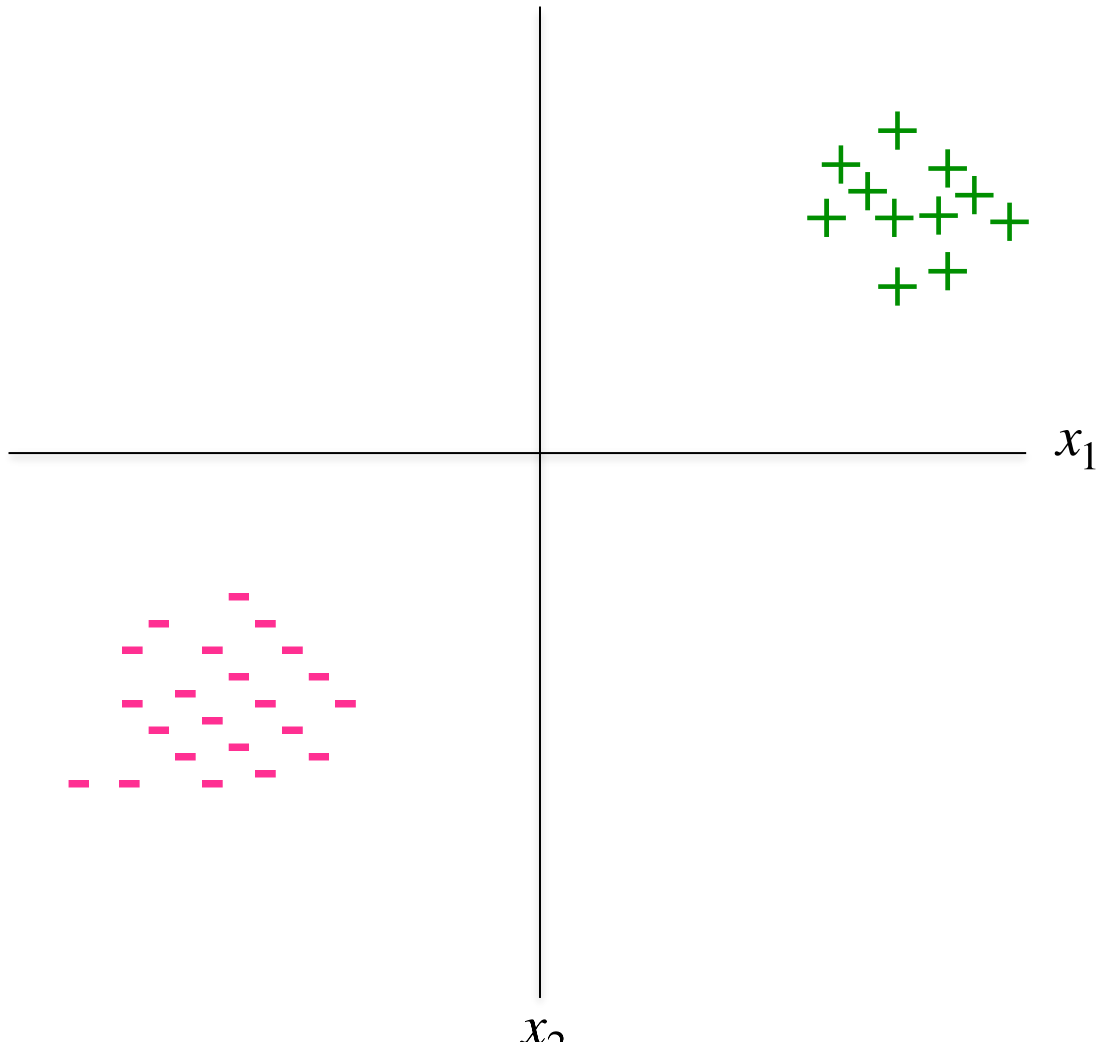
In 2 dimensions: length $= ||\mathbf{v}|| = \sqrt{v_1^2 + v_2^2}$

In 3 dimensions: length $= ||\mathbf{v}|| = \sqrt{v_1^2 + v_2^2 + v_e^2}$

See Introduction to Linear Algebra by Gilbert Strang

The geometry of a linear classifier

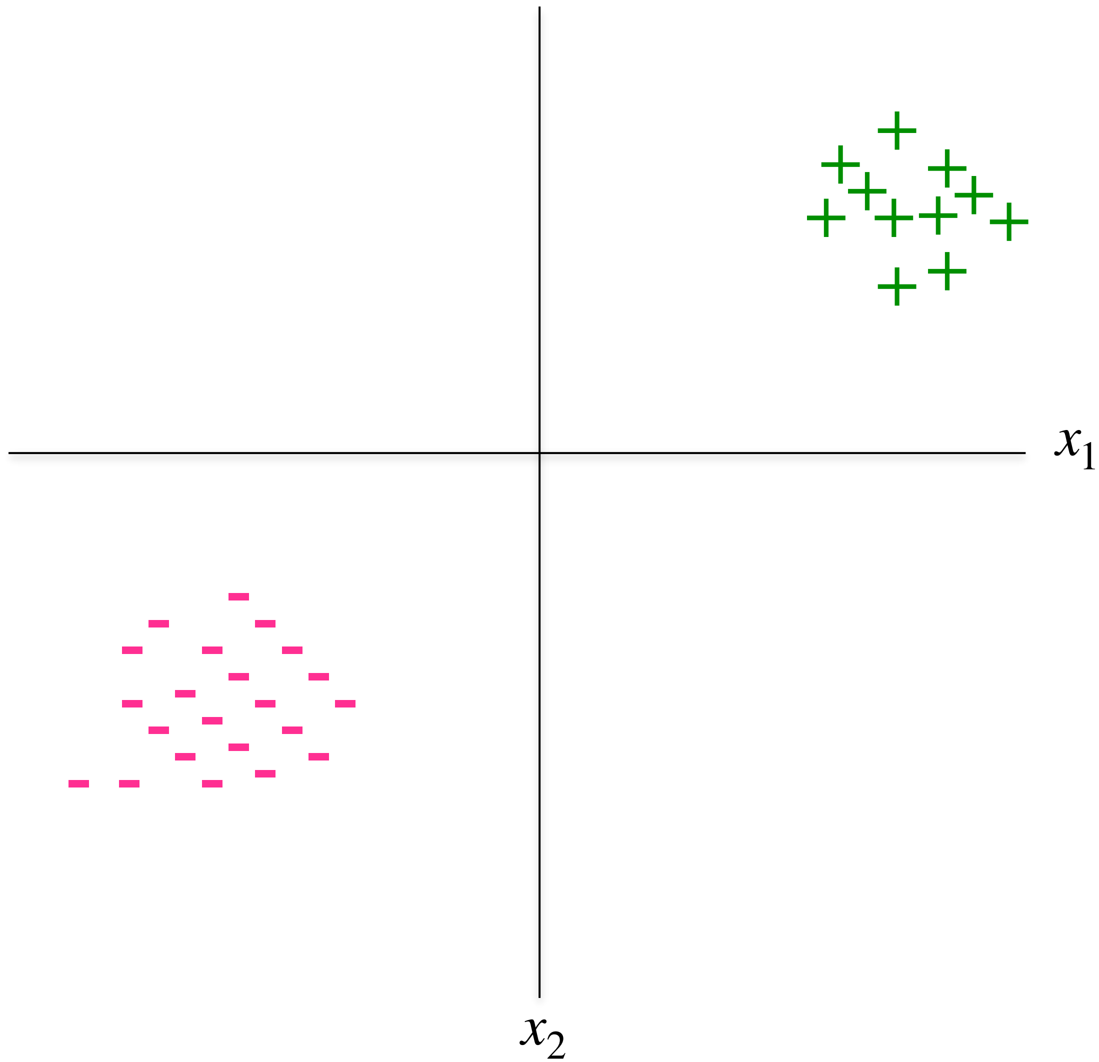
An illustration in two dimensions



The geometry of a linear classifier

An illustration in two dimensions

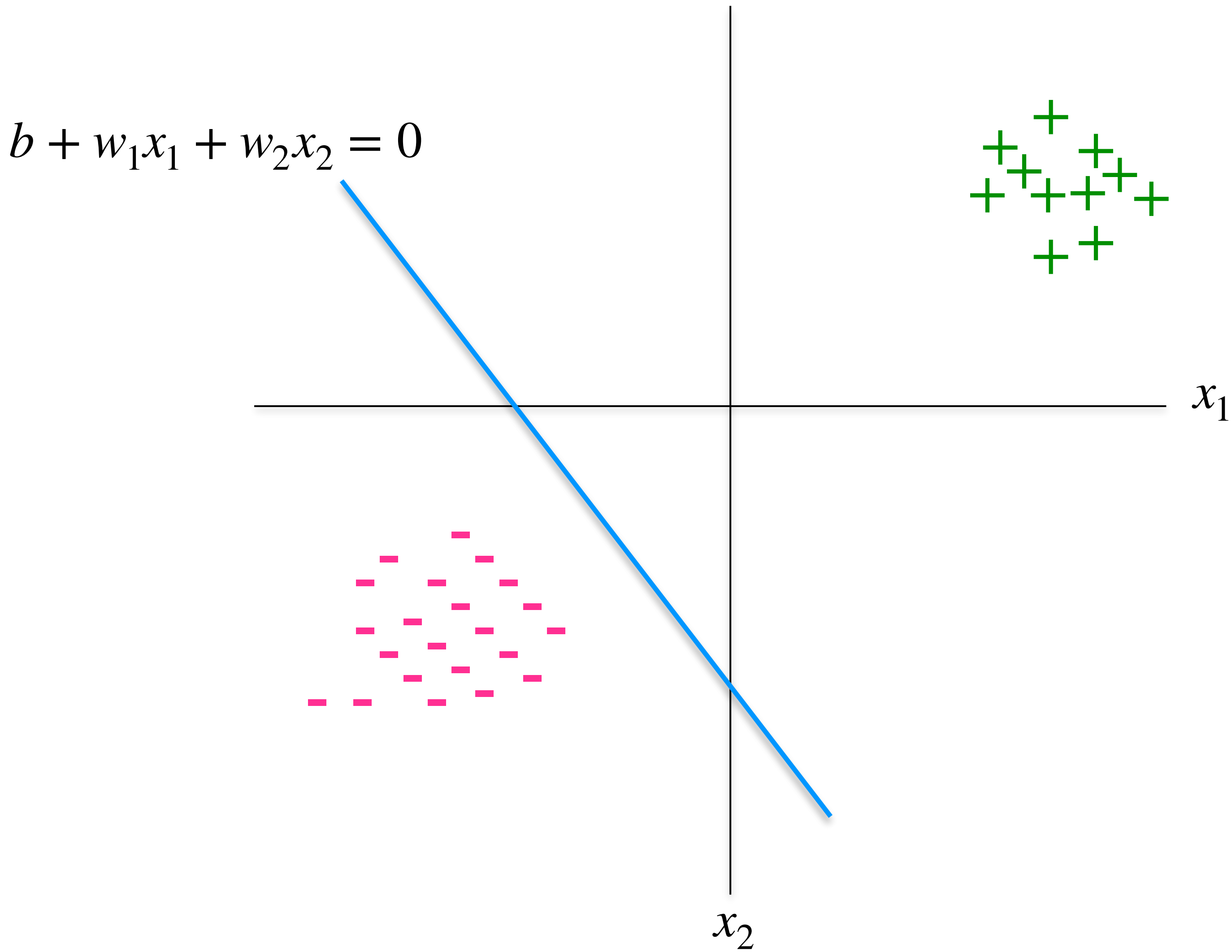
$\text{sgn}(b + w_1 x_1 + w_2 x_2)$



The geometry of a linear classifier

An illustration in two dimensions

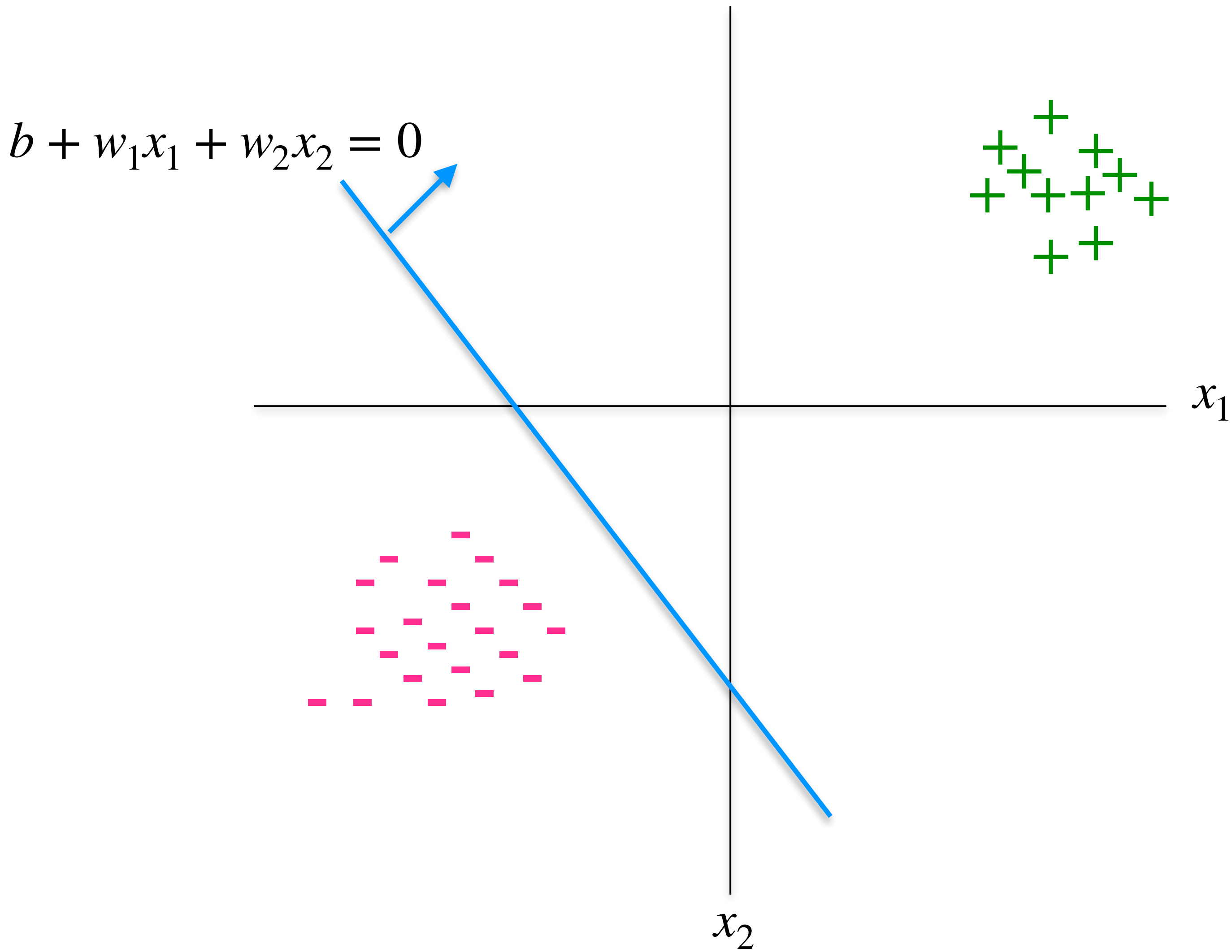
$$\text{sgn}(b + w_1 x_1 + w_2 x_2)$$



The geometry of a linear classifier

An illustration in two dimensions

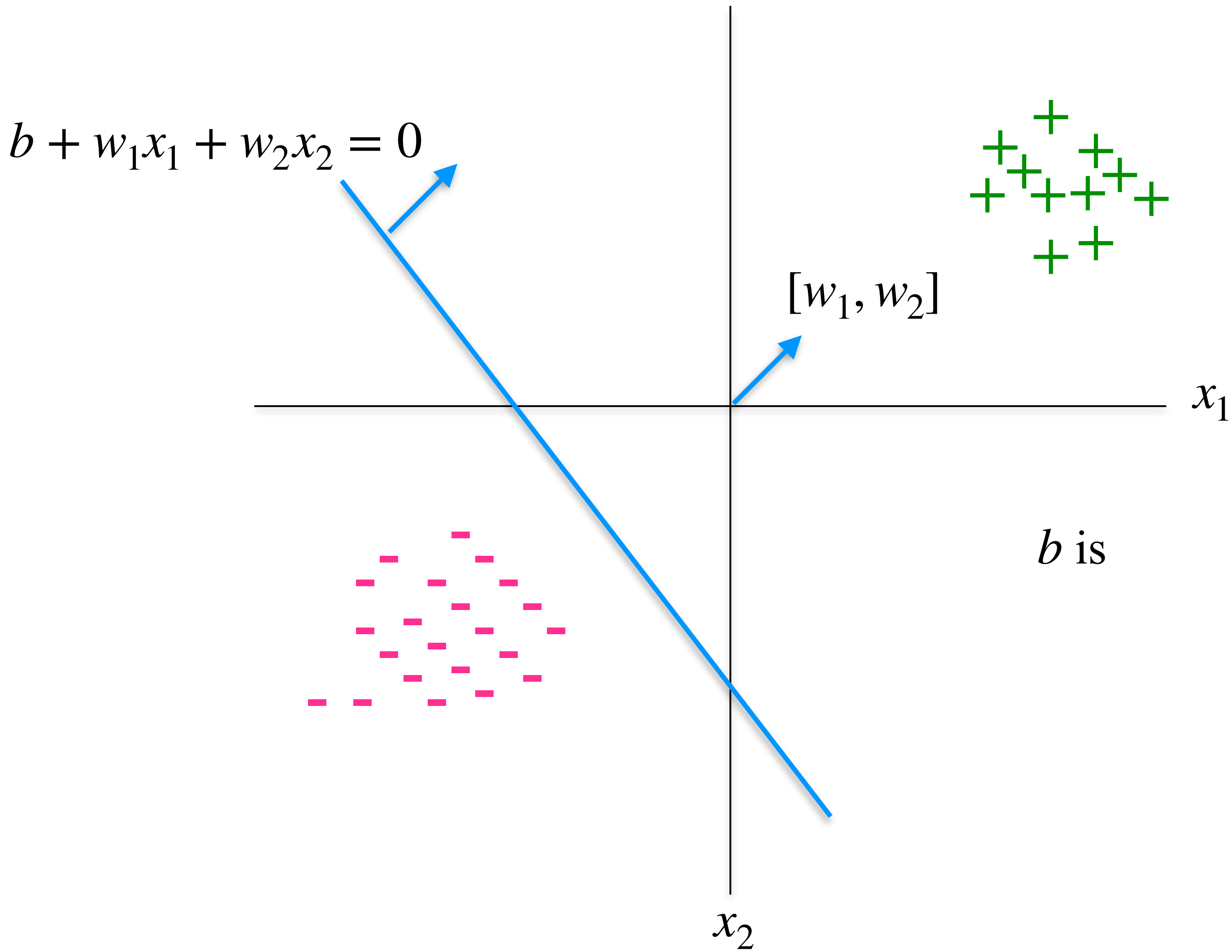
$$\text{sgn}(b + w_1 x_1 + w_2 x_2)$$



The geometry of a linear classifier

An illustration in two dimensions

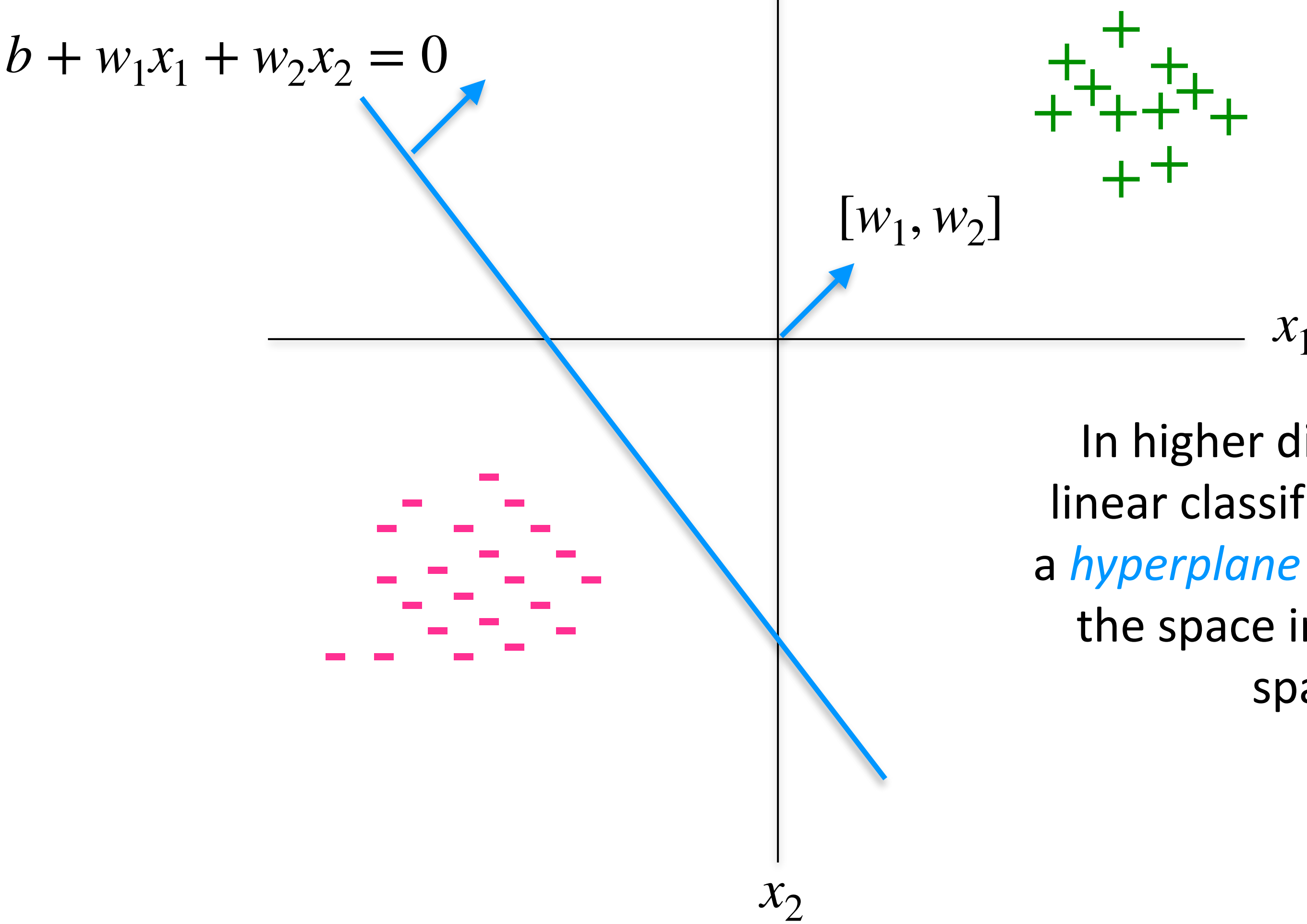
$\text{sgn}(b + w_1x_1 + w_2x_2)$



The geometry of a linear classifier

$$\mathrm{sgn}(b + w_1 x_1 + w_2 x_2)$$

We only care about the sign, not the magnitude



In higher dimensions, a linear classifier represents a *hyperplane* that separates the space into two half-spaces

Simplifying notation

We can stop writing b at each step using notational sugar:

The prediction function is $\text{sgn}(\mathbf{w}^T\mathbf{x} + b) = \text{sgn}(\sum_i w_i x_i + b)$

Rewrite $\mathbf{x}$ as $\begin{bmatrix}\mathbf{x}\\1\end{bmatrix}$. Call this $\mathbf{x}'$. Rewrite $\mathbf{w}$ as $\begin{bmatrix}\mathbf{w}\\b\end{bmatrix}$. Call this $\mathbf{w}'$

Note that $\mathbf{w}^T\mathbf{x} + b$ is the same as $\mathbf{w}'^T\mathbf{x}'$

The prediction function is now $\text{sgn}(\mathbf{w}'^T\mathbf{x}')$

Increases dimensionality by one

Equivalent to adding a feature that is constant: always 1

In the increased dimensional space, the vector w' goes through the origin

We sometimes hide the bias b, and instead fold the bias term into the weights by adding an extra constant feature. But remember that it is there.

Coming up: linear classification

Perceptron: error driven learning, updates the hypothesis if there is an error

Logistic regression: another probabilistic classifier

Naive Bayes classifier: a simple linear classifier with a probabilistic interpretation

In all cases, the prediction will be done with the same rule:

$$\mathbf{w}^T\mathbf{x} + b \geq 0 \Rightarrow y = +1$$
$$\mathbf{w}^T\mathbf{x} + b < 0 \Rightarrow y = -1$$

Linear Classifiers

EXPRESSIVENESS

Where are we?

Linear models: introduction

What functions do linear classifiers express?

- Conjunctions and disjunctions
- m-of-n functions
- Not all functions are linearly separable
- Feature space transformations
- Exercises

Which Boolean functions can linear classifiers represent?

Linear classifiers are an expressive hypothesis class

Many Boolean functions are linearly separable

- Not all though
- Recall: In comparison, decision trees can represent any Boolean function

Conjunctions and disjunctions

$y = x_1 \wedge x_2 \wedge x_3$ is equivalent to "$y = 1$ whenever $x_1 + x_2 + x_3 \geq 3$"

x_1	x_2	x_3	y	$x_1 + x_2 + x_3 = 3$	sign
0	0	0	0	-3	0
0	0	1	0	-2	0
0	1	0	0	-2	0
0	1	1	0	-1	0
1	0	0	0	-2	0
1	0	1	0	-1	0
1	1	0	0	-1	0
1	1	1	1	0	1

Negations are okay too. In general, use $1 - x$ in the linear threshold unit if x is negated

$y = x_1 \wedge x_2 \wedge \neg x_3$ corresponds to

$$x_1 \wedge x_2 \wedge (1 - x_3) \geq 3$$

m-of-n functions

m-of-n rules

- There is a fixed set of n variables
- $y =$ true if and only if at least m of them are true
- All other variables are ignored

Suppose there are five Boolean variables: x_1, x_2, x_3, x_4, x_5

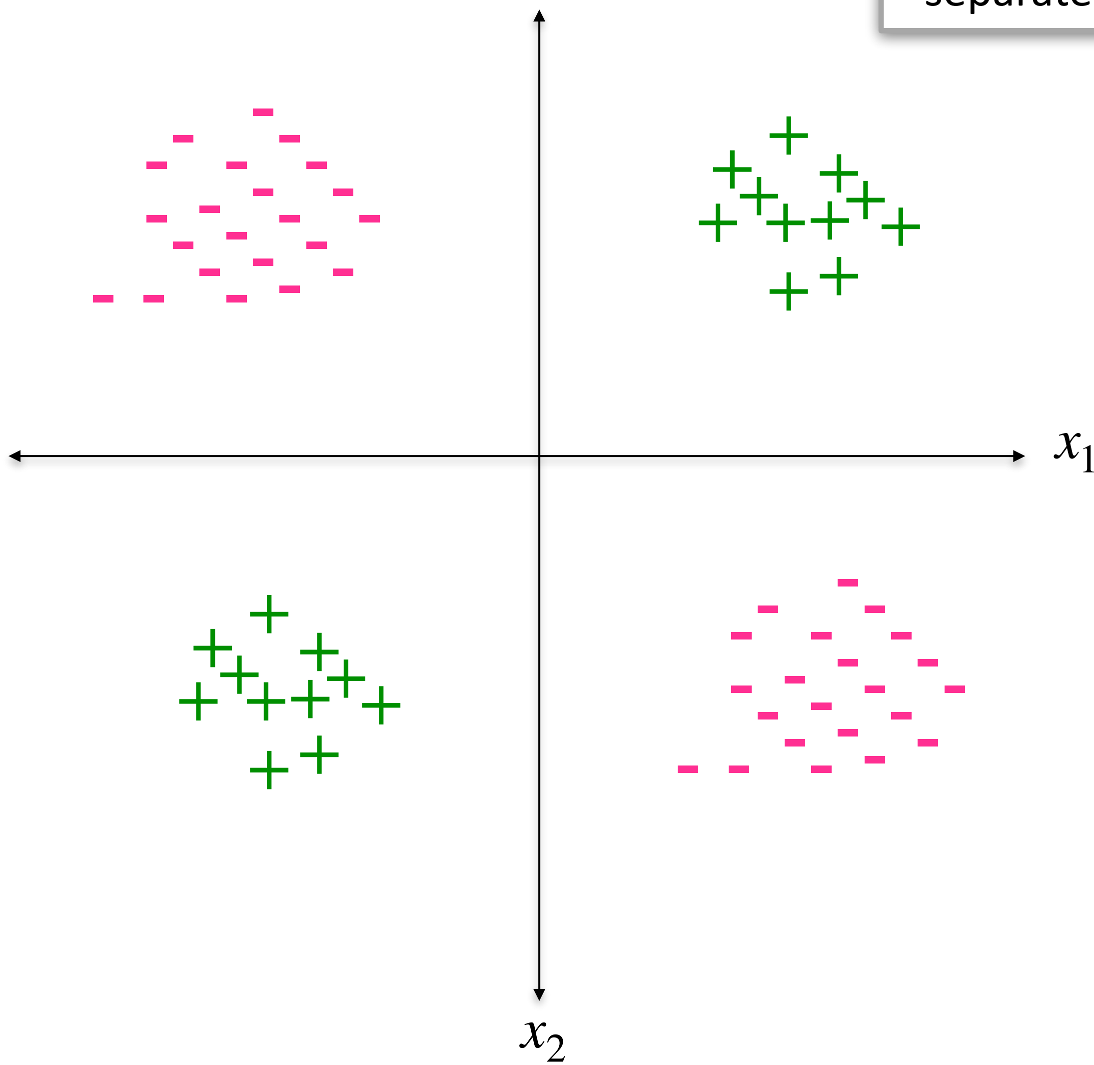
What is a threshold unit that is equivalent to the classification rule "at least 2 of $\{x_1, x_2, x_3\}$"?

$$x_1 + x_2 + x_3 \geq 2$$

Parity is not linearly separable

(The XOR function)

Can't draw a line to separate the two classes



Not all functions are linearly separable

XOR is not linear

- $y = x$ XOR y
- $y = (x \land \neg y) \lor (\neg x \land y)$
- Parity cannot be represented as a linear classifiers
  - $f(\mathbf{x}) = 1$ if the number of 1s is even

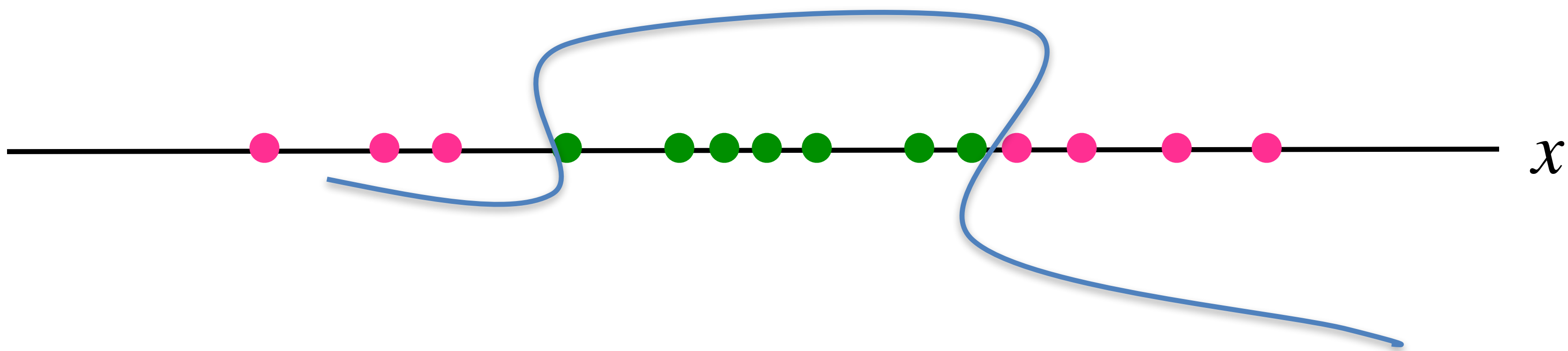
Many non-trivial Boolean functions

- Example: $y = (x_1 \land x_2) \lor (x_3 \land \neg x_4)$
- The function is not linear in the four variables

Even these functions can be made linear

These points are not separable in 1-dimension by a line

What is a one-dimensional line, by the way?

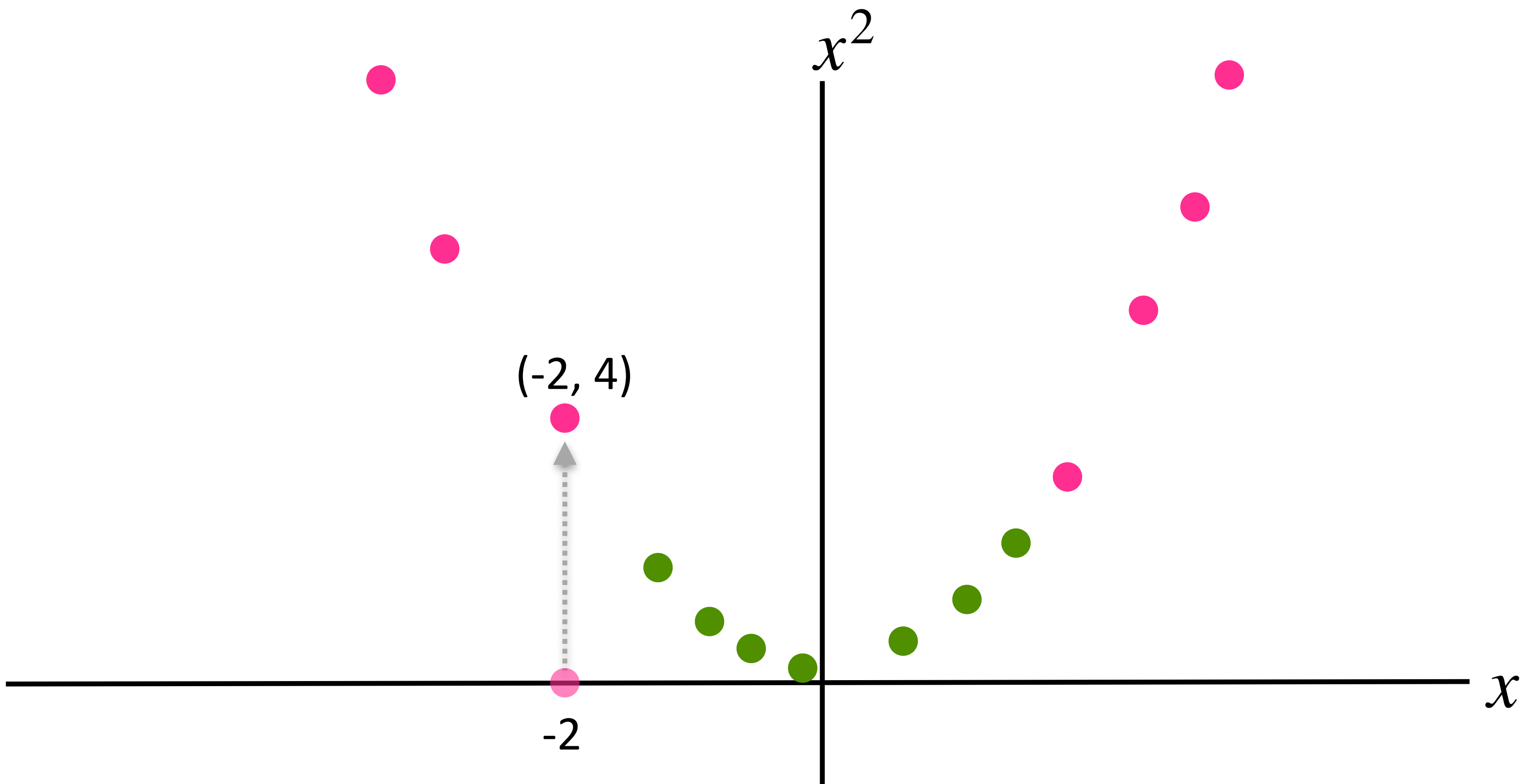


The trick: change the representation

The blown up feature space

The trick: use feature conjunctions

Transform points: represent each point x in 2 dimensions by (x, x^2)

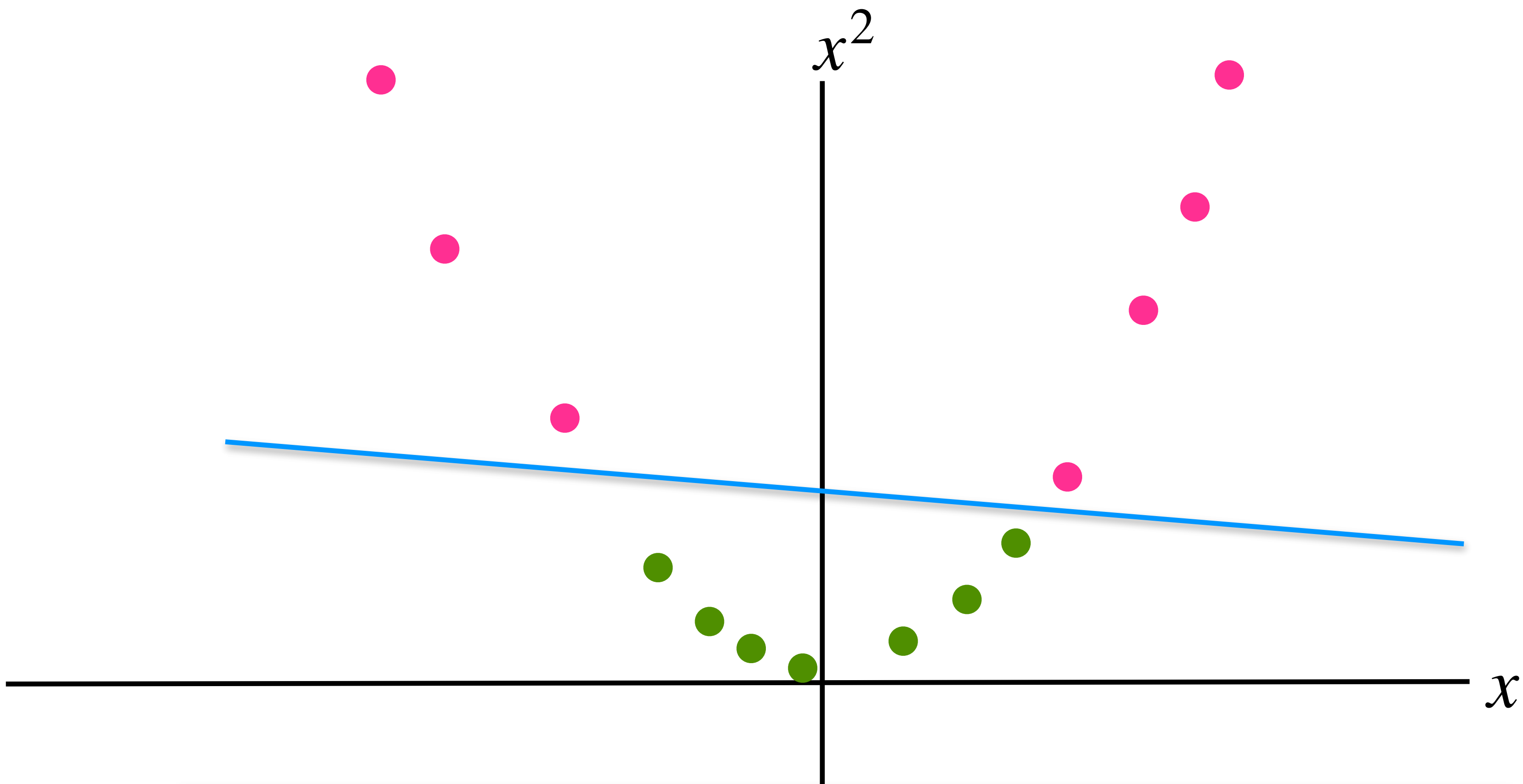


Now the data is linearly separable in this space!

The blown up feature space

The trick: use feature conjunctions

Transform points: represent each point x in 2 dimensions by (x, x^2)



Key issue: representation. What features to use?

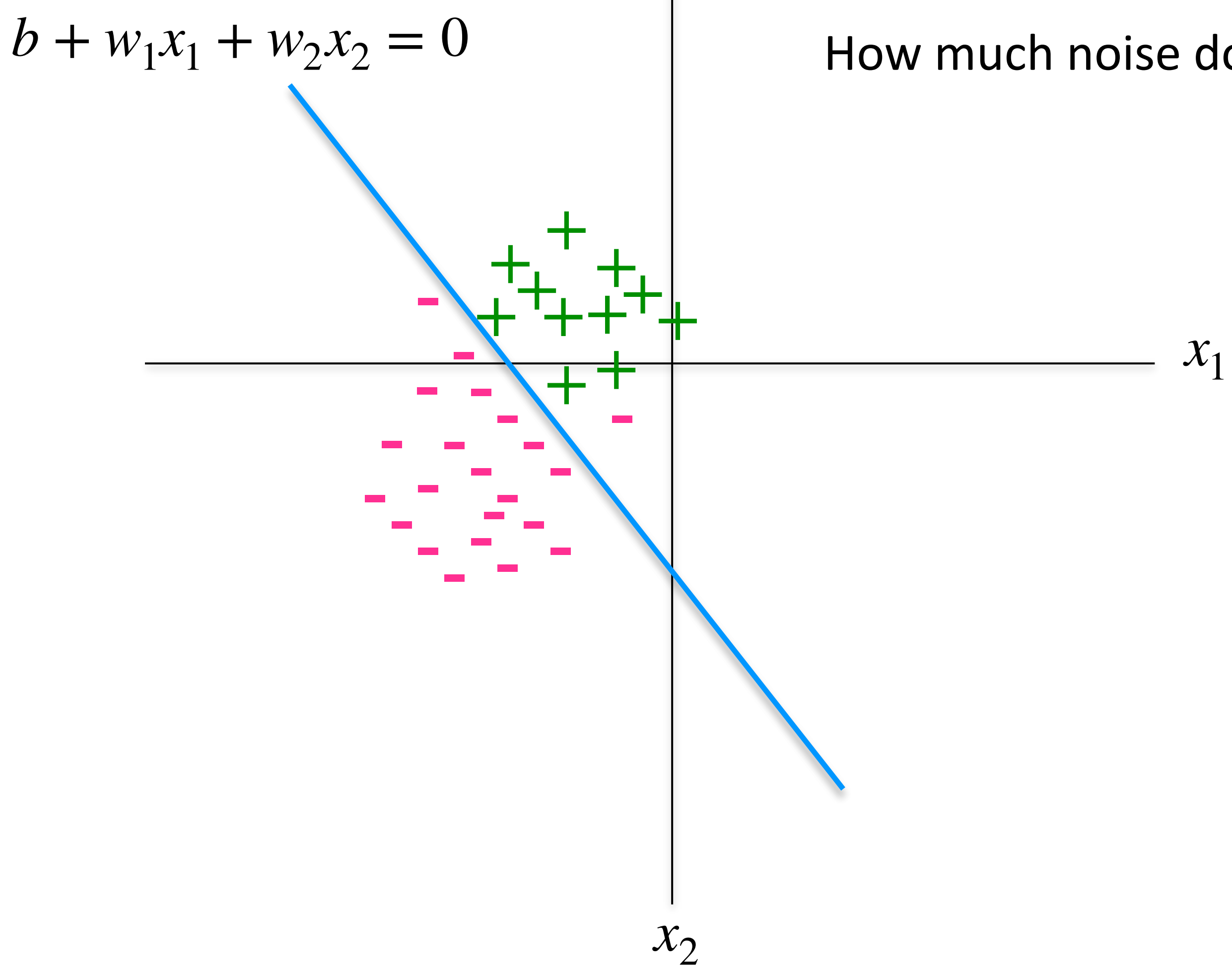
Exercise

How would you use the feature transformation idea to make XOR in two dimensions in two dimensions linearly separable in a new space?

To answer this question, you need to think about a function that maps examples from two dimensional space to a higher dimensional space.

Almost linearly separable data

$\text{sgn}(b + w_1 x_1 + w_2 x_2)$



Training data is almost separable, except for some noise

How much noise do we allow for?

Linear classifiers: an expressive hypothesis class

Many functions are linear

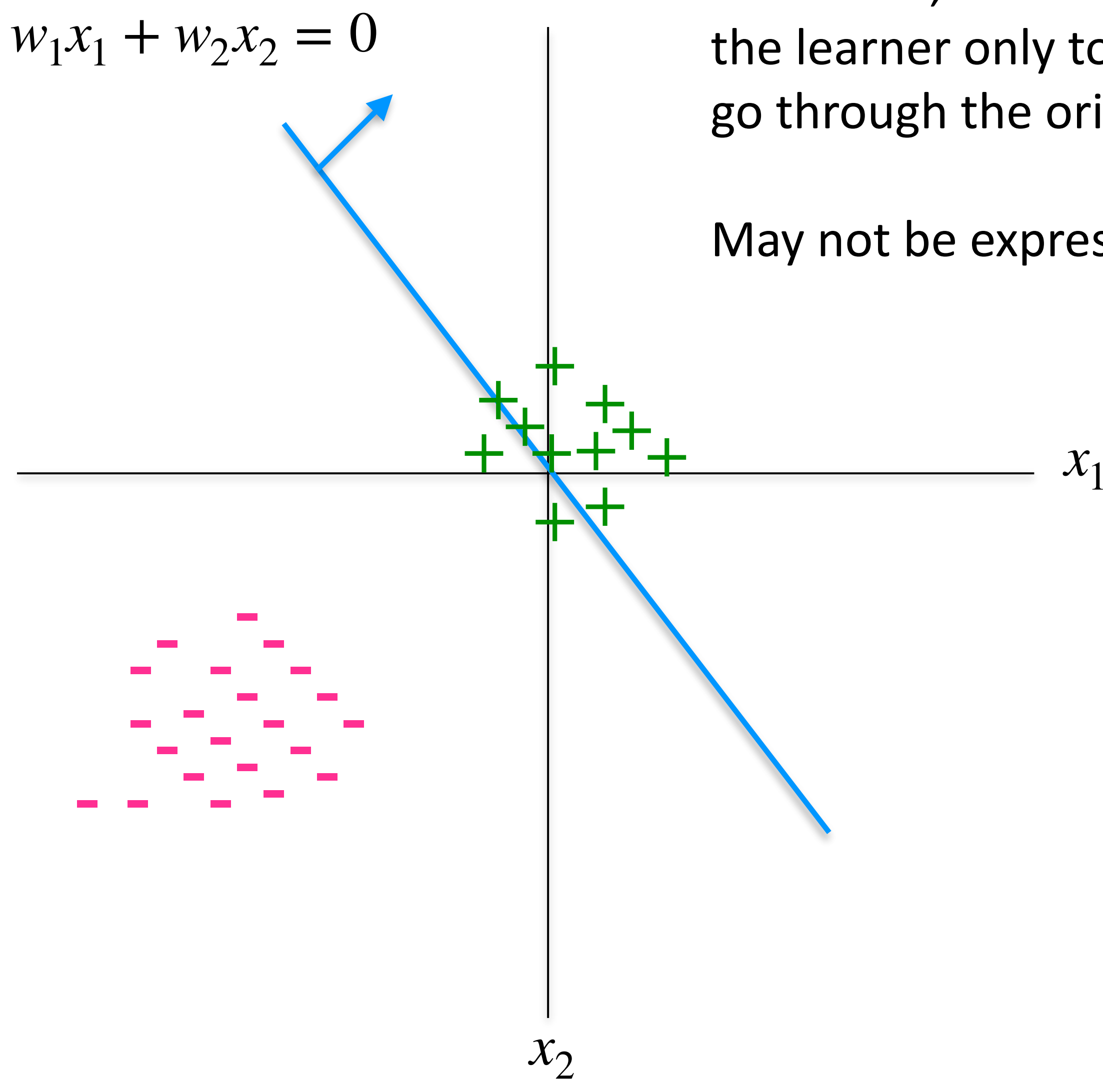
Often a good guess for a hypothesis space

Some functions are not linear

- The XOR function
- Non-trivial Boolean functions

But there are ways of making them linear in a higher dimensional feature space

Why is the bias term needed?



If b is zero, then we are restricting the learner only to hyperplane that go through the origin

May not be expressive enough