LINEAR CLASSIFICATION: INTRODUCTION

Guy Azran & Itay Evron

Tutorial outline

- Classification (recap)
- Linear classification
  - Separability
  - Linear separators & classification
  - Higher dimensions
  - Non-homogenous separation

Classification

- Recall:
  - Features: $\boldsymbol{x} \in \mathbb{R}^d$
  - Labels: $y \in \{\pm 1\}$
  - Data dist.: $(\boldsymbol{x}, y) \overset{iid}{\sim} D = D_{XY}$

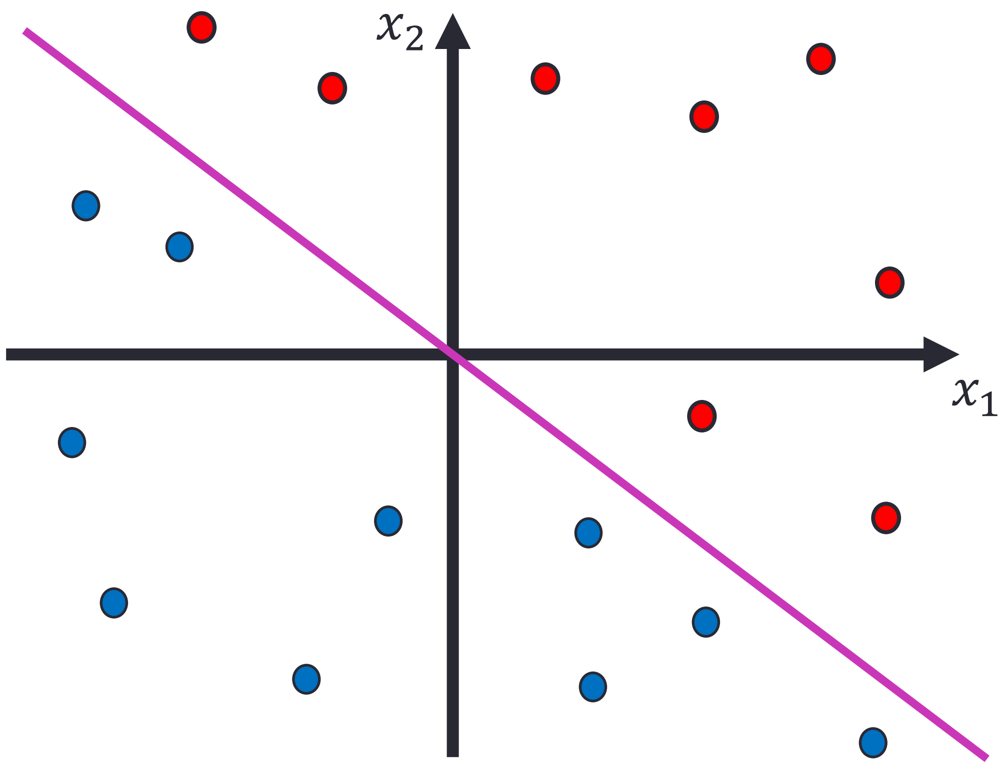
- Consider an example $\boldsymbol{x} \in \mathbb{R}^d$ with an unknown binary label $y \in \{\pm 1\}$.
- A binary classifier, or hypothesis, $h: \mathbb{R}^d \to \{\pm 1\}$, predicts $\hat{y} = h(\boldsymbol{x})$.

- In classification tasks, we aim to find a classifier h that is correct ($\hat{y} = y$) with high probability.

LINEAR CLASSIFICATION

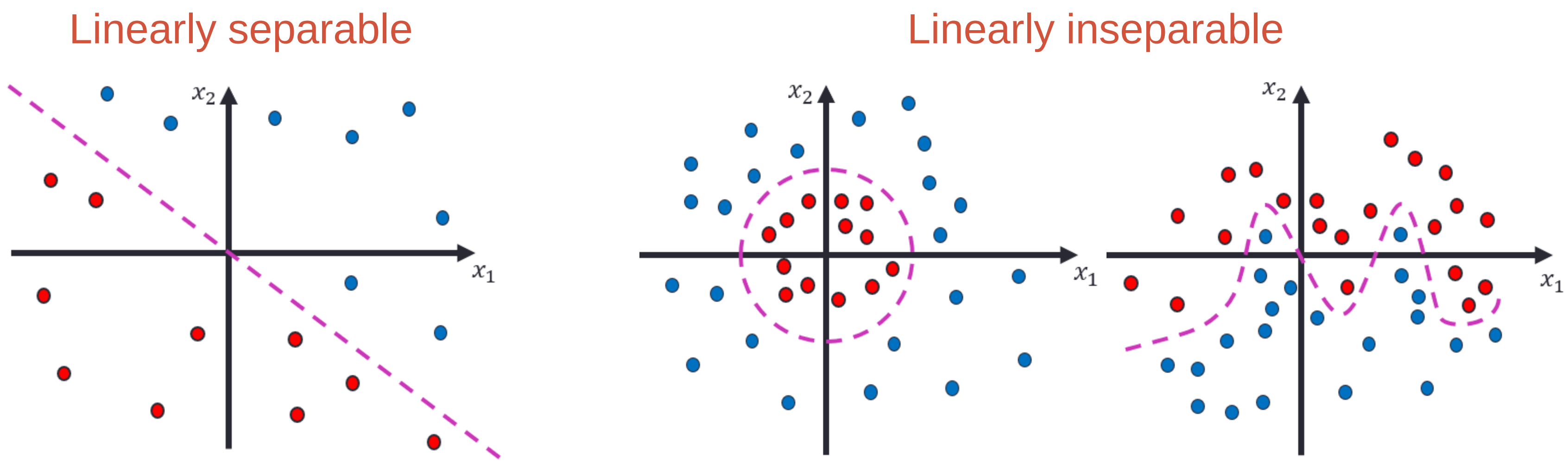
Linear classification

- Very important problems in machine learning.
- We will investigate linear models throughout the course.
- Today: develop intuition using linear algebra.



Separability

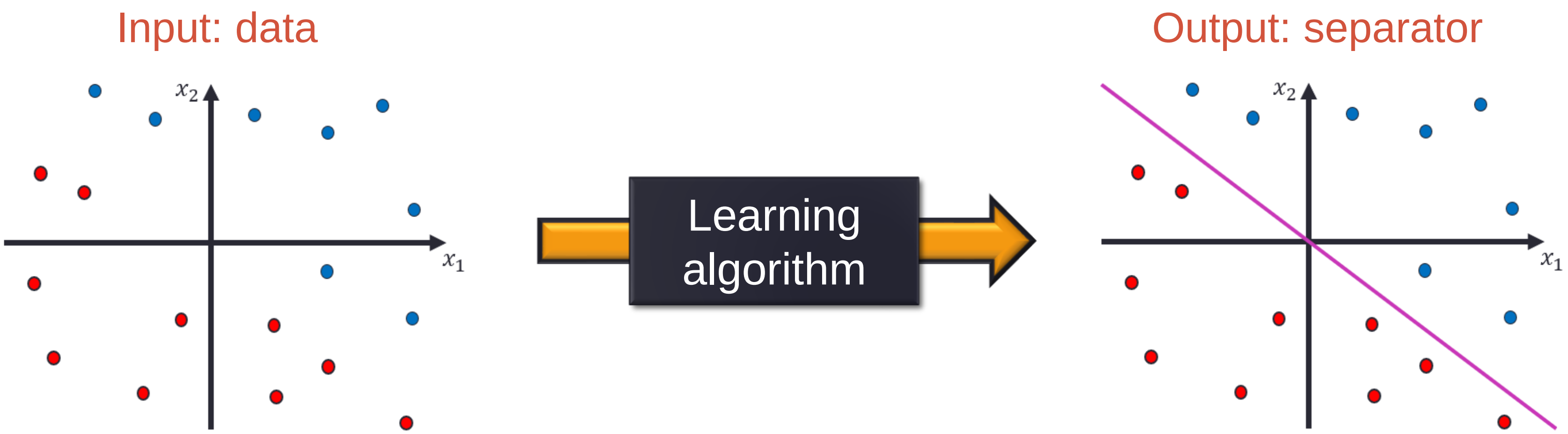
- If a dataset can be classified correctly using a linear separator, it is called linearly separable.



- For now, we concentrate on linearly-separable cases.

The separator

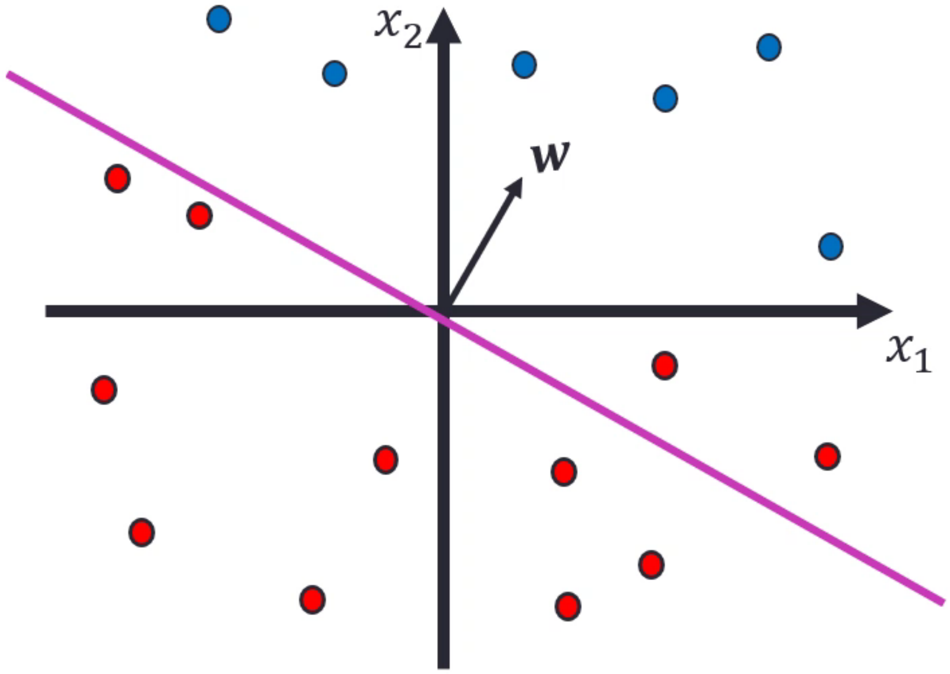
- Assume a black-box capable of finding a separator (classifier).



- Question: what should the output of the black box be?
  - Line equation (m,b)?
  - Two points on the line?

Normal vector

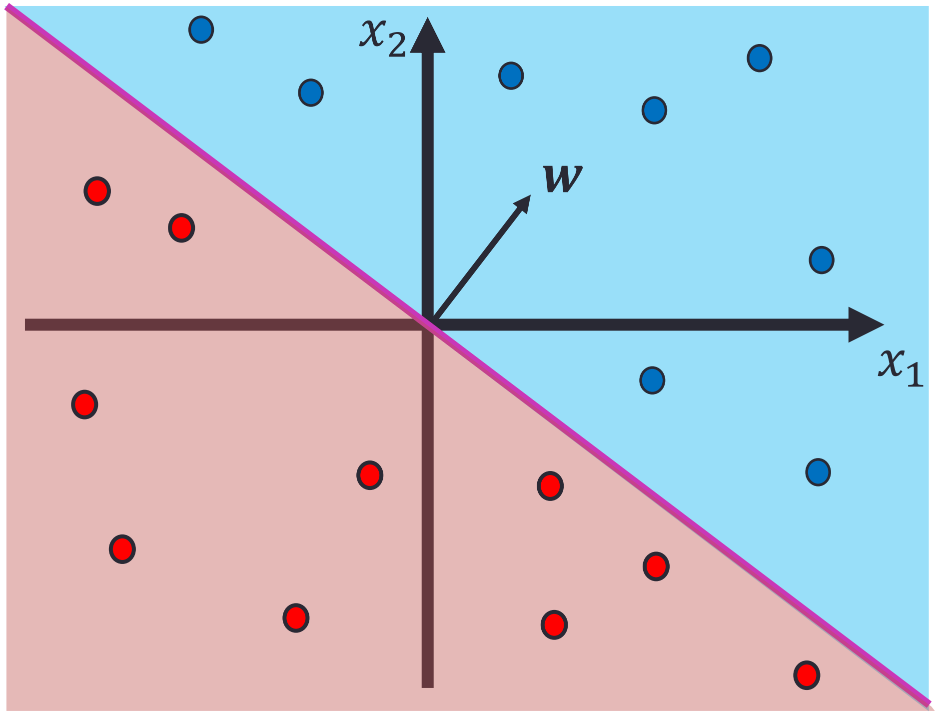
- Define a separating line by using a normal $\boldsymbol{w}$ perpendicular to it.
- Compact representation: a single vector.
- Generalizes well to higher dimensions.



Algebraic intuition

- Consider our toy dataset in $\mathbb{R}^2$.
- The black-box produced $\boldsymbol{w}$.
- Exercise: design a function

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \textit{ is blue} \\ -1, & \boldsymbol{x} \textit{ is red} \end{cases}$$

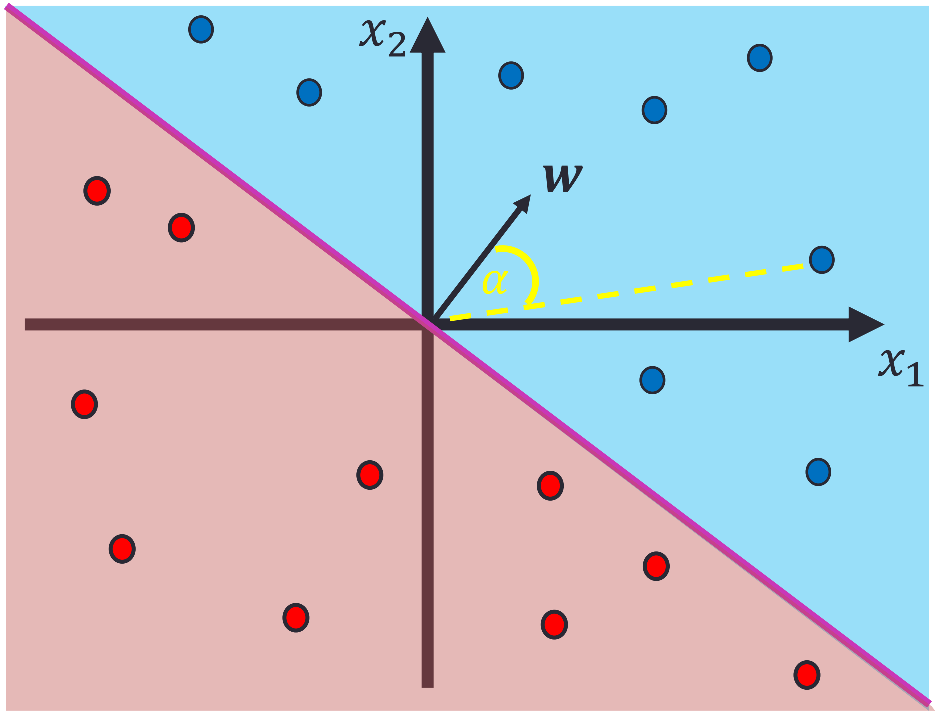


Algebraic intuition

- Consider our toy dataset in $\mathbb{R}^2$.
- The black-box produced $\boldsymbol{w}$.
- Exercise: design a function

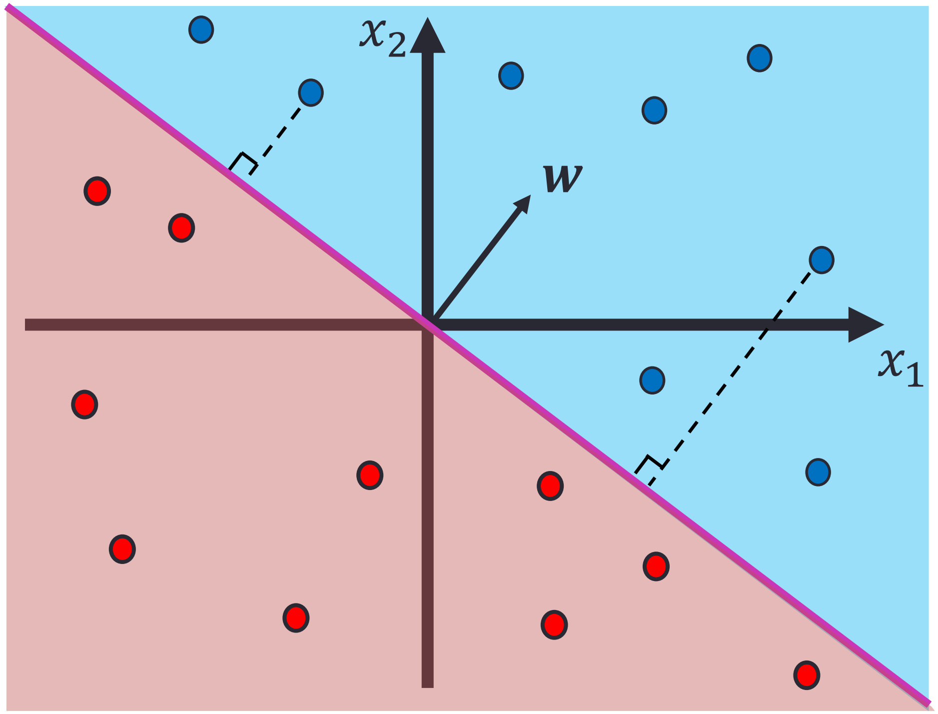
$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \textit{ is blue} \\ -1, & \boldsymbol{x} \textit{ is red} \end{cases}$$

- Reminder: $\boldsymbol{w}^\top \boldsymbol{x} = \|\boldsymbol{w}\| \|\boldsymbol{x}\| \cos \alpha$.
- What is the sign of this value?



Margin

- Are all blue data points born equal, or are some “bluer” than others?
- The distance between data points and the separator indicates confidence.
- This distance is called the margin.
- Show: the margin is $\boldsymbol{w}^\top \boldsymbol{x} / \|\boldsymbol{w}\|$.

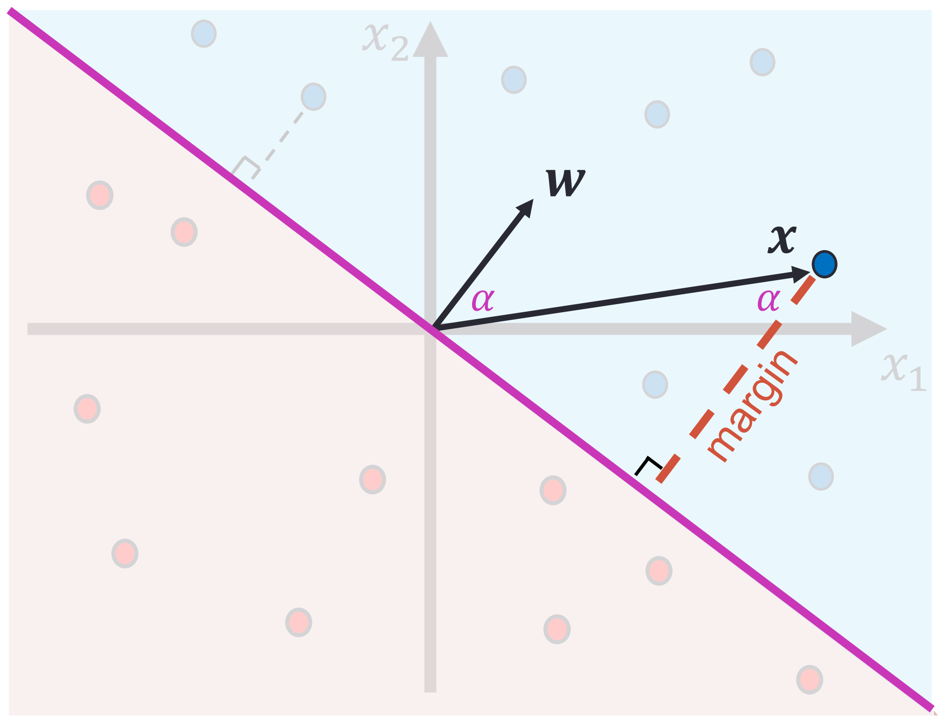


Margin

- Are all blue data points born equal, or are some "bluer" than others?
- The distance between data points and the separator indicates confidence.
- This distance is called the margin.
- Show: the margin is $\boldsymbol{w}^\top \boldsymbol{x} / \|\boldsymbol{w}\|$.
- Remember: $\boldsymbol{w}^\top \boldsymbol{x} = \|\boldsymbol{w}\| \|\boldsymbol{x}\| \cos \alpha$.
- Basic trigonometry: $\cos \alpha = \frac{\text{margin}}{\|\boldsymbol{x}\|}$

$$\Longrightarrow \text{margin} = \frac{\boldsymbol{w}^\top \boldsymbol{x}}{\|\boldsymbol{w}\|} \quad \blacksquare$$

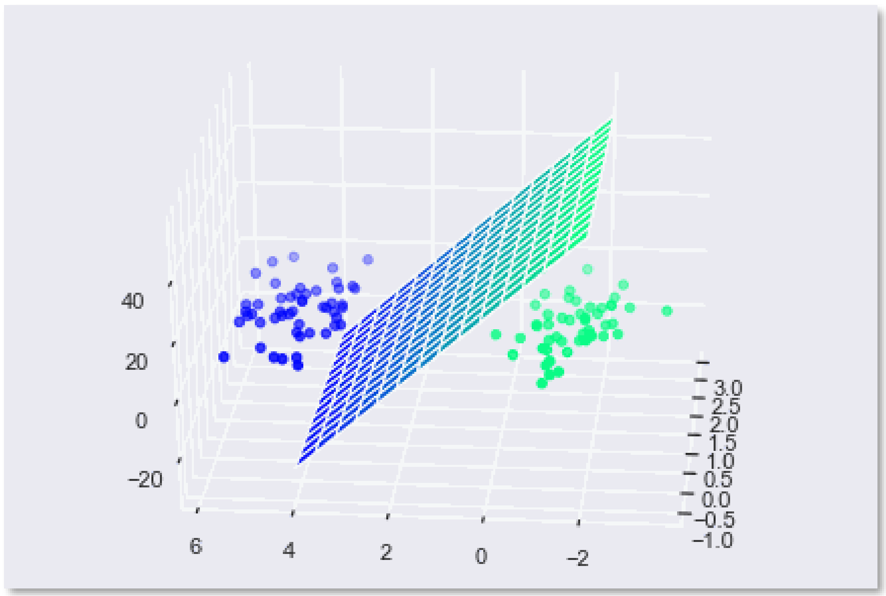
- Note: it is negative when $\alpha > \frac{\pi}{2}$



Higher dimensions

"To deal with hyperplanes in a 14-dimensional space, visualize a 3d space and say "fourteen" to yourself very loudly. Everyone does it."

- Geoffrey Hinton



Source: MLJ

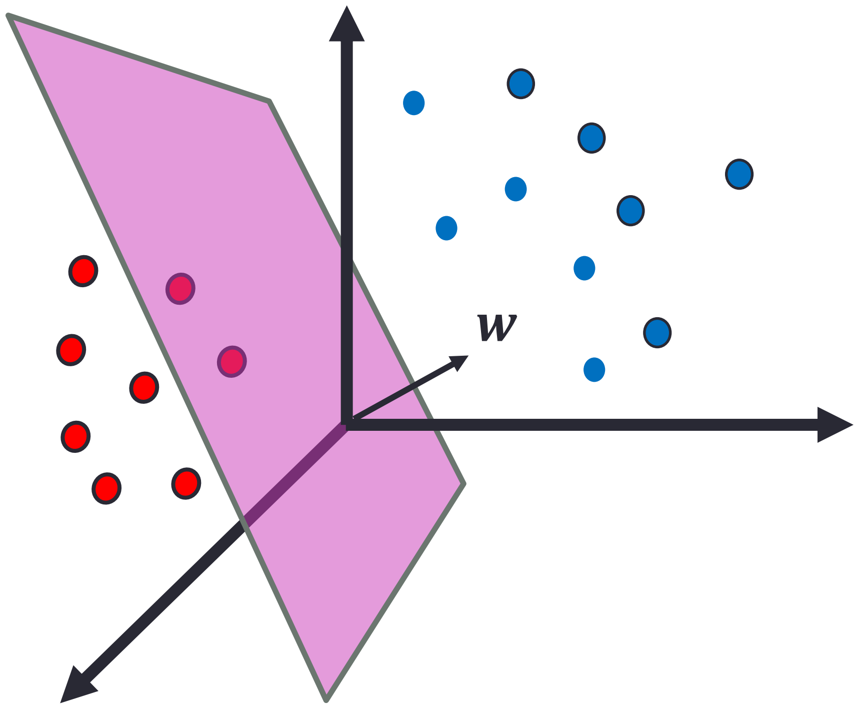
- Instead of a separating line we have a separating hyperplane.

Higher dimensions

- The algebra we used in 2d is still valid.
- Our black-box will yield a normal $\boldsymbol{w}$, perpendicular to a separating hyperplane.
- Use the same classifier:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^\top \boldsymbol{x})$$

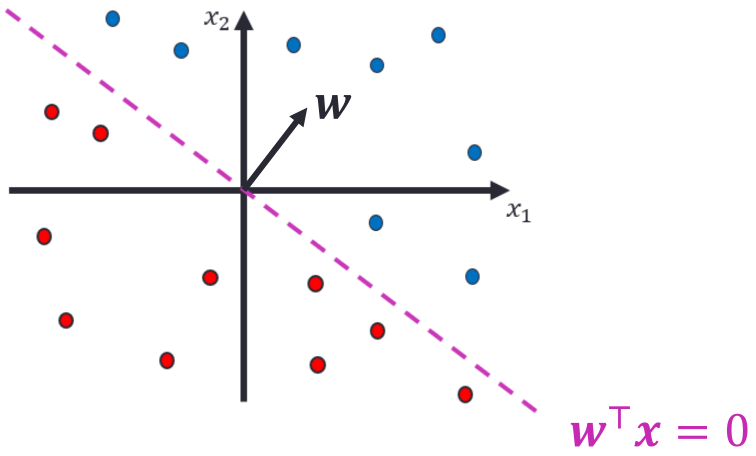
- In this course,
  we will learn how to find good
  separating hyperplanes.



Decision boundary

- The linear classifiers we saw take the form $h_{\boldsymbol{w}}(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^\top \boldsymbol{x})$.
- Q: mathematically, where is the decision boundary?
  - A: the set of all data points where $\boldsymbol{w}^\top \boldsymbol{x} = 0$.

Homogeneous linear separator

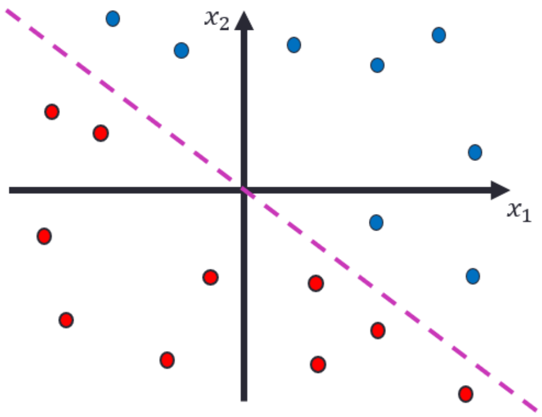


This is a homogeneous linear equation!

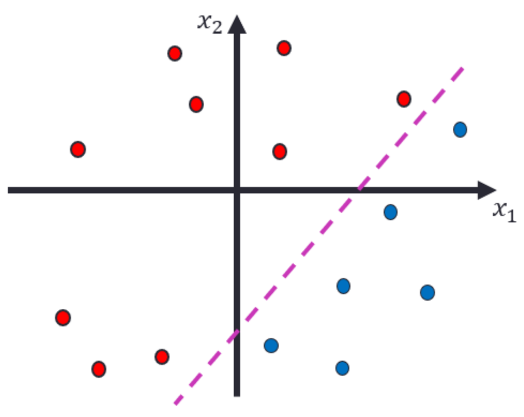
Non-homogeneous linear separation

- What if the data is not centered?
- Q: what changes in the hypothesis class?

Homogeneous
linear separator



Non-homogeneous
linear separator



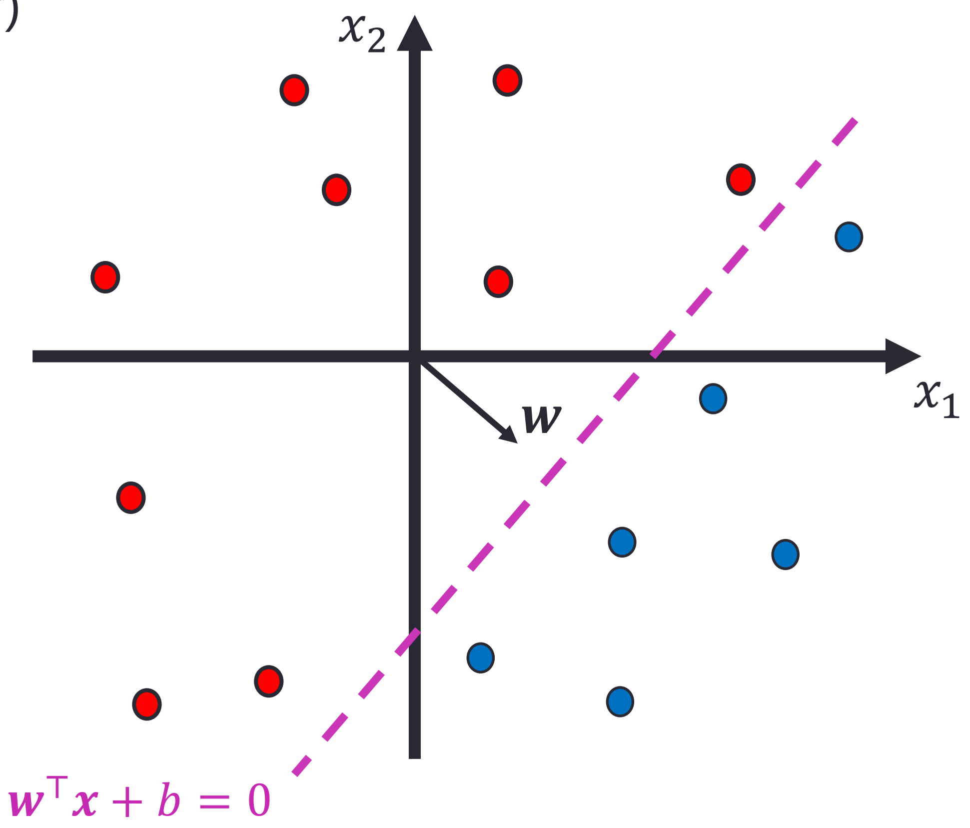
Non-homogeneous linear separation

- What if the data is not centered?
- Q: what changes in the hypothesis class?
  - A: simply add a bias term (scalar)
- The decision rule:

$$h_{\boldsymbol{w},b}(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b)$$

- The decision boundary:

$$\boldsymbol{w}^\top \boldsymbol{x} + b = 0$$



Extension to non-homogeneous

- Exercise: extend our black-box to non-homogeneous cases.

$$h_{\boldsymbol{w},b}(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b)$$

- Reduction to homogeneous:
  - Add a constant feature to all examples
  - Find a $(d+1)$-dimensional homogeneous separator

$$\mathrm{sign}(\boldsymbol{w}^\top \boldsymbol{x} + b) = \mathrm{sign}\left(\begin{bmatrix}\boldsymbol{w}\\ b\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{x}\\ 1\end{bmatrix}\right)$$

- We can extend every homogeneous linear model like that!

Exercise: linearly dependent features

- Consider a 2d training set: $\mathbf{X} \in \mathbb{R}^{m\times 2}, \boldsymbol{y} \in \mathbb{R}^{m}$.
- Assume: The best training accuracy achievable by a linear classifier is 90%.
- We add a 3rd feature by summing the 2 original features.
  - For instance, $\boldsymbol{x}_i = [x_{i,1},\ x_{i,2}]$ transforms into $\boldsymbol{x}_i' = [x_{i,1},\ x_{i,2},\ \ x_{i,1} + x_{i,2}]$.

- Question: the best training accuracy achievable by a linear classifier is now:

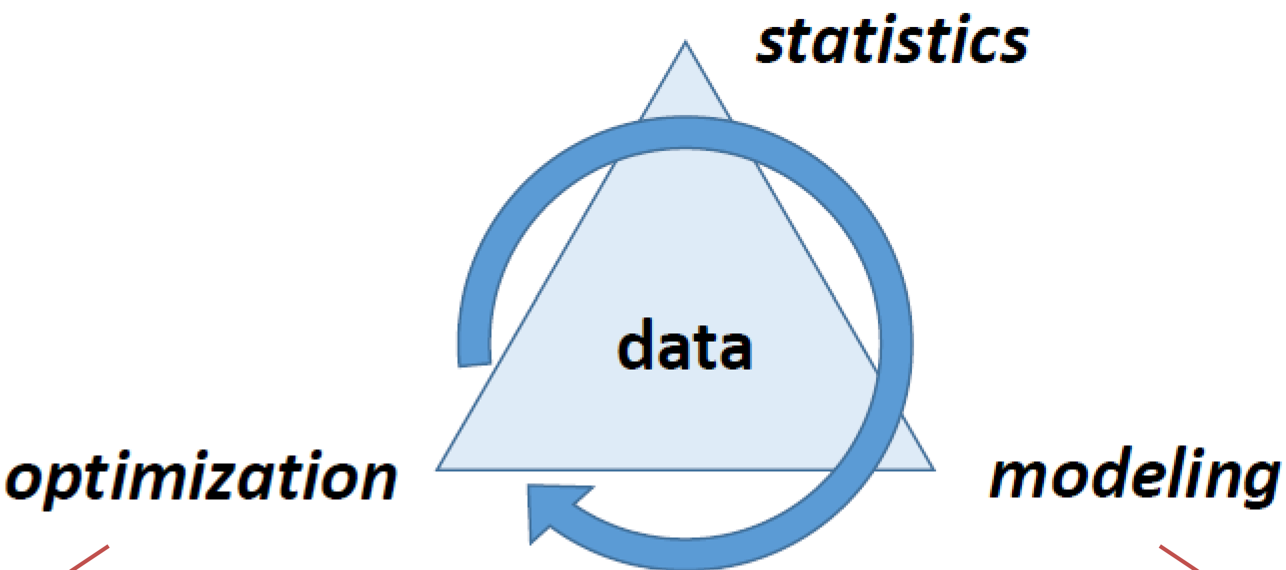
  (a) Higher (≥) (b) Lower (≤) (c) Unchanged (=)

- Extra: what happens if we transform the sample to $[x_{1,1},\ x_{1,2},\ \ x_{1,1}^2]$?

Three pillars of learning

- We will study linear classification from different perspectives.
- Which perspective did we use so far?

How much data do we need?
Week 5



Find a “good” hypothesis efficiently
Weeks 3, 4, 7, 8

Define a hypothesis class
Today, week 3