ARTIFICIAL INTELLIGENCE 2

Support Vector Machines

Francesco Spinnato

* francesco.spinnato@unipi.it
University of Pisa

Introduction

We are in the usual classification scenario:

- Each sample: $\mathbf{x}_i \in \mathbb{R}^M$
- Each label: $y_i \in \{-1, +1\}$

Goal: learn a function that correctly classifies new samples.

What's the best decision boundary?

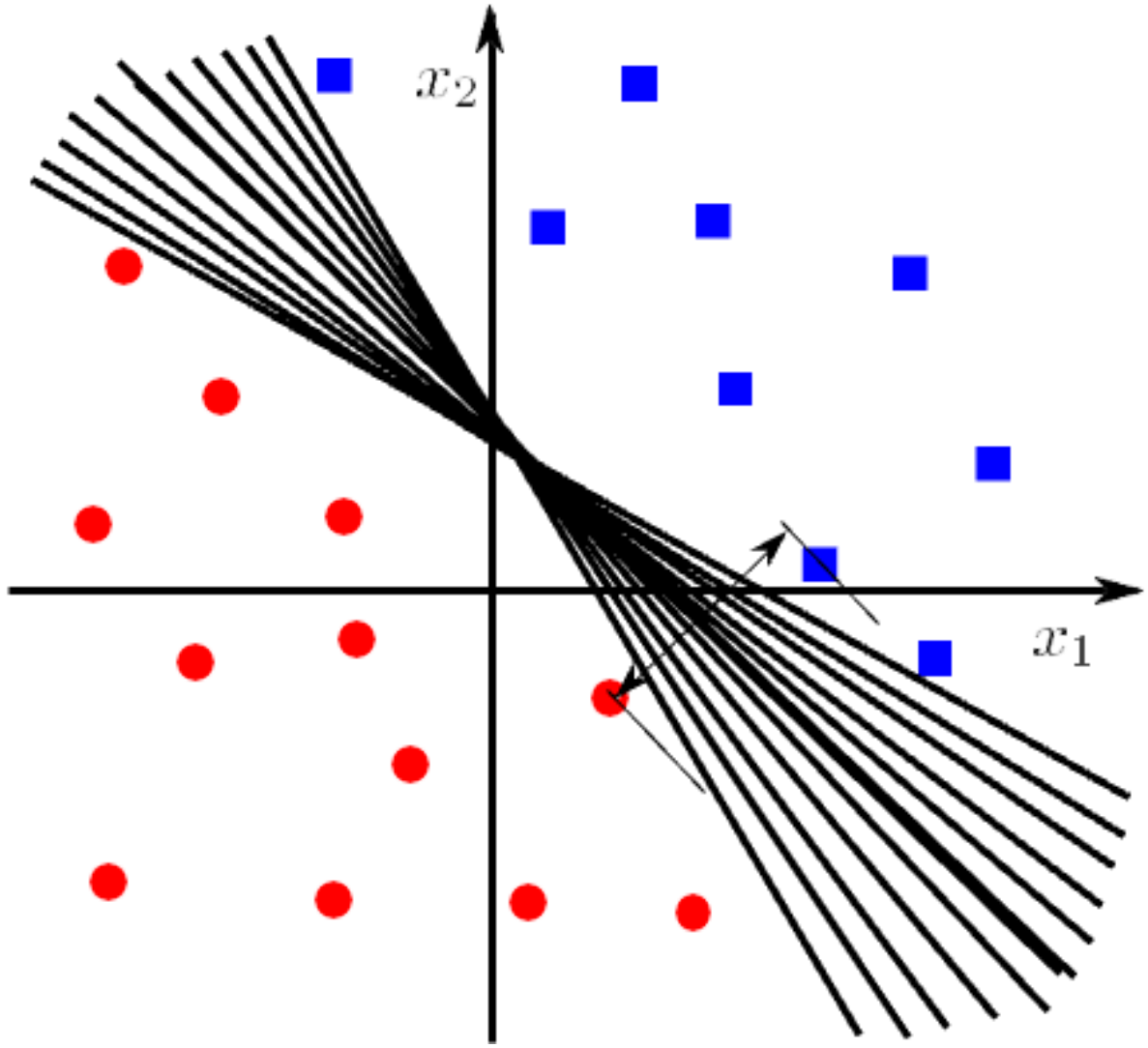
A linear classifier separates classes with a hyperplane (a line in 2D):

$$\sum_{j=0}^{M} w_j x_j = 0$$

with the prediction rule being:

$$\text{sign}\left(\sum_{j=0}^{M} w_j x_j\right)$$

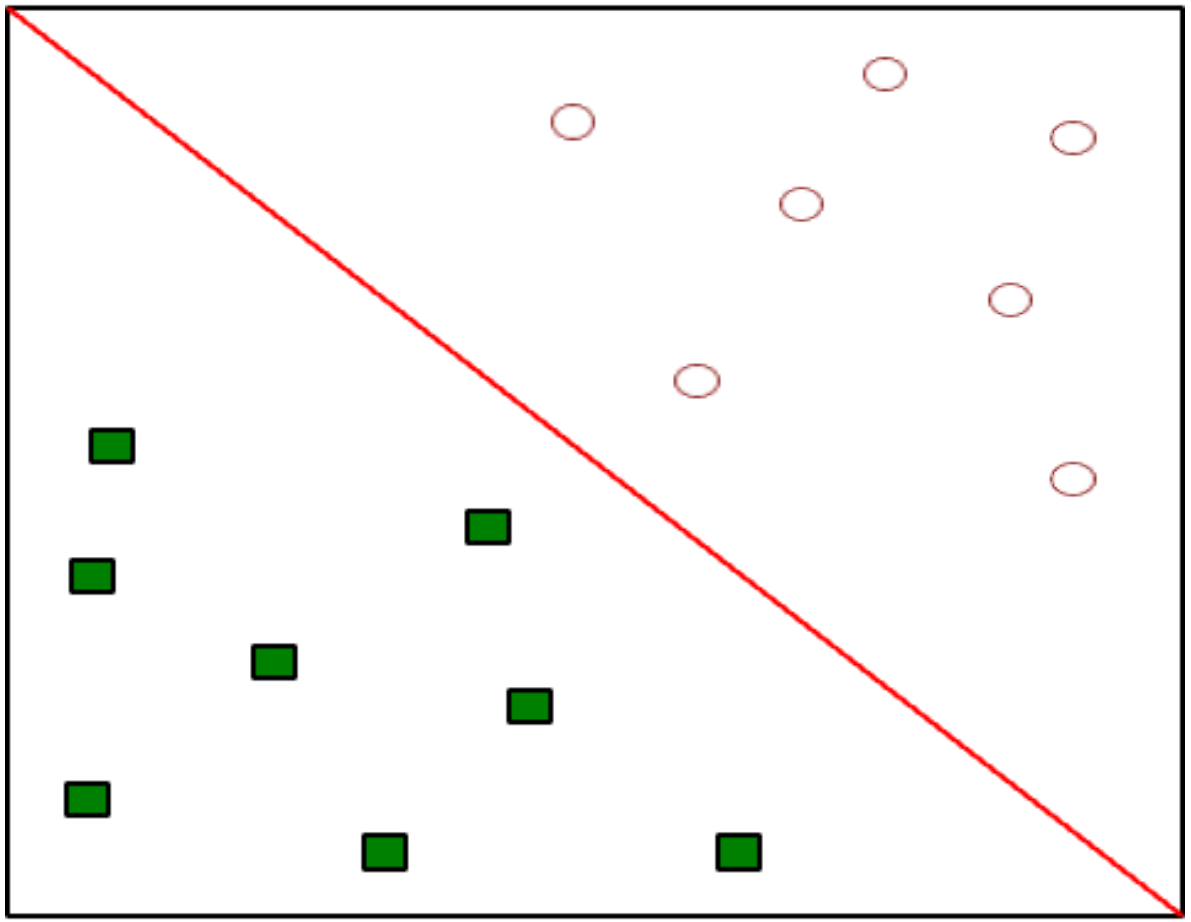
Given a linearly separable set of points, there are **infinite hyperplanes** that could solve the classification problem.



*Assuming a dummy 0-th column containing all ones, so w_0 is the bias

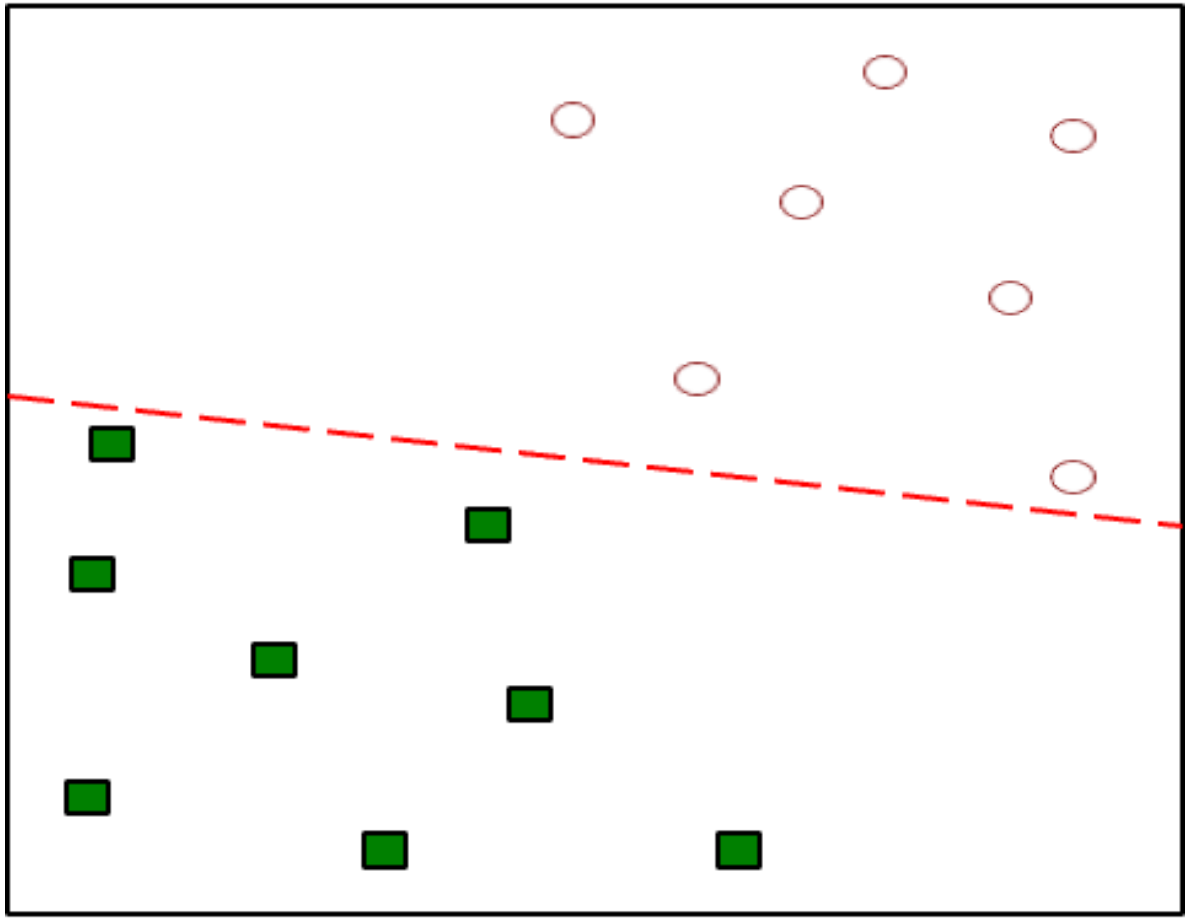
What's the best decision boundary?

This is one possible solution



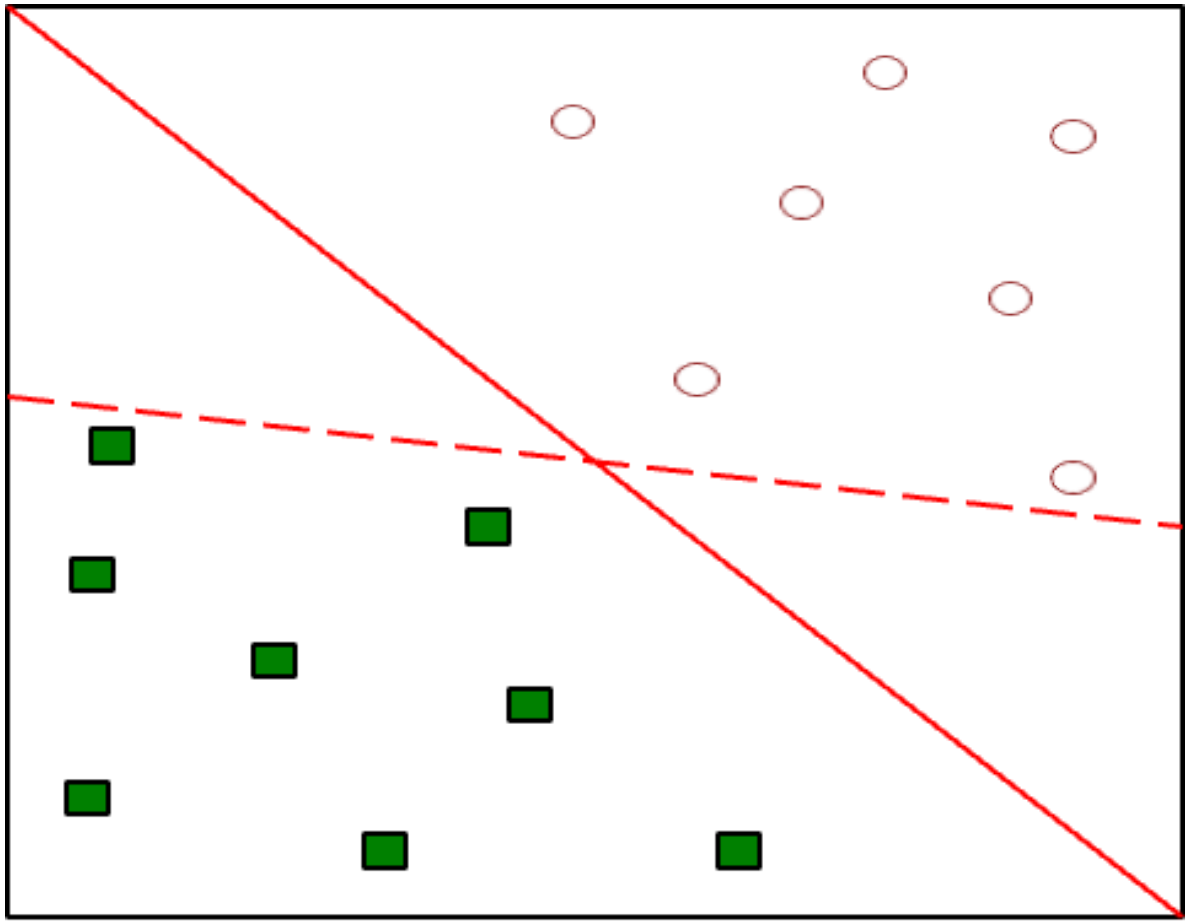
What's the best decision boundary?

But this is also valid



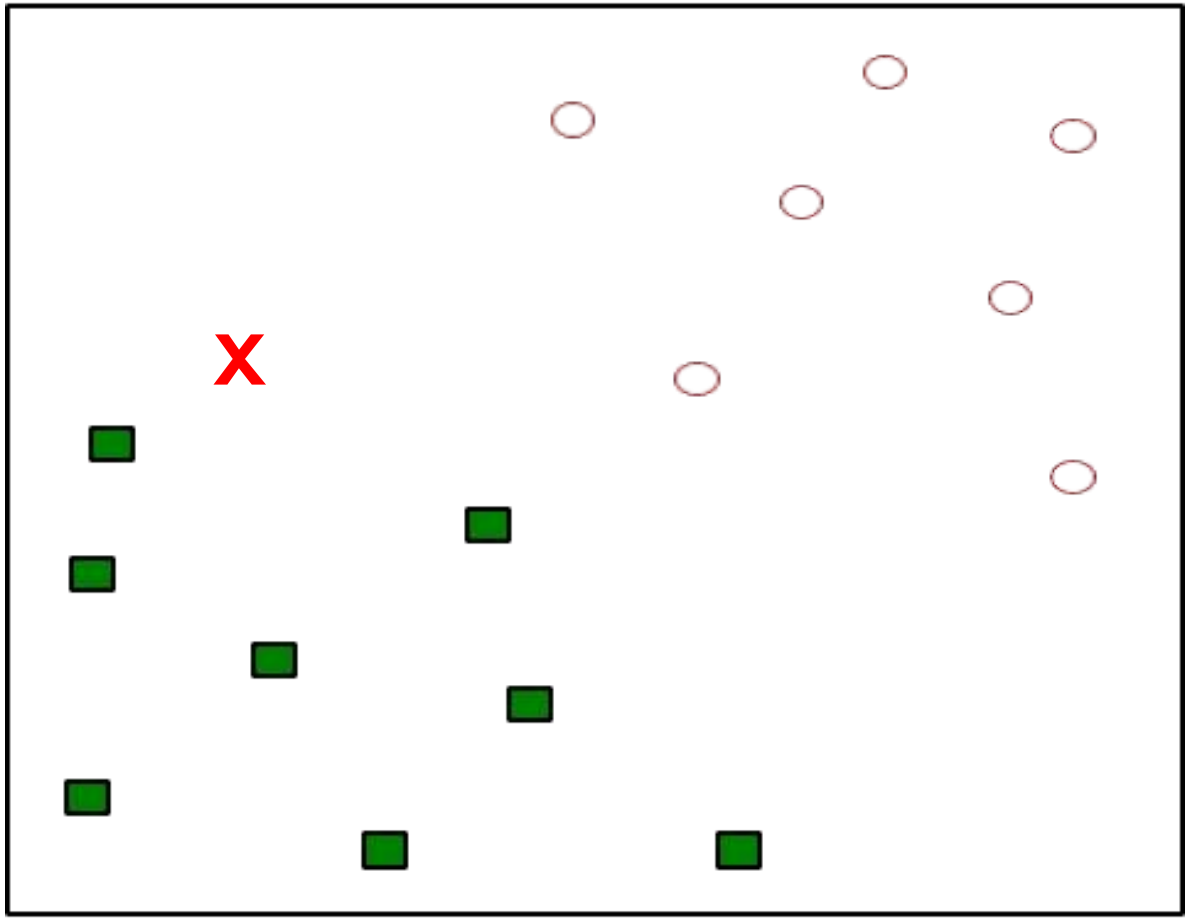
What's the best decision boundary?

How do you decide which one is better?



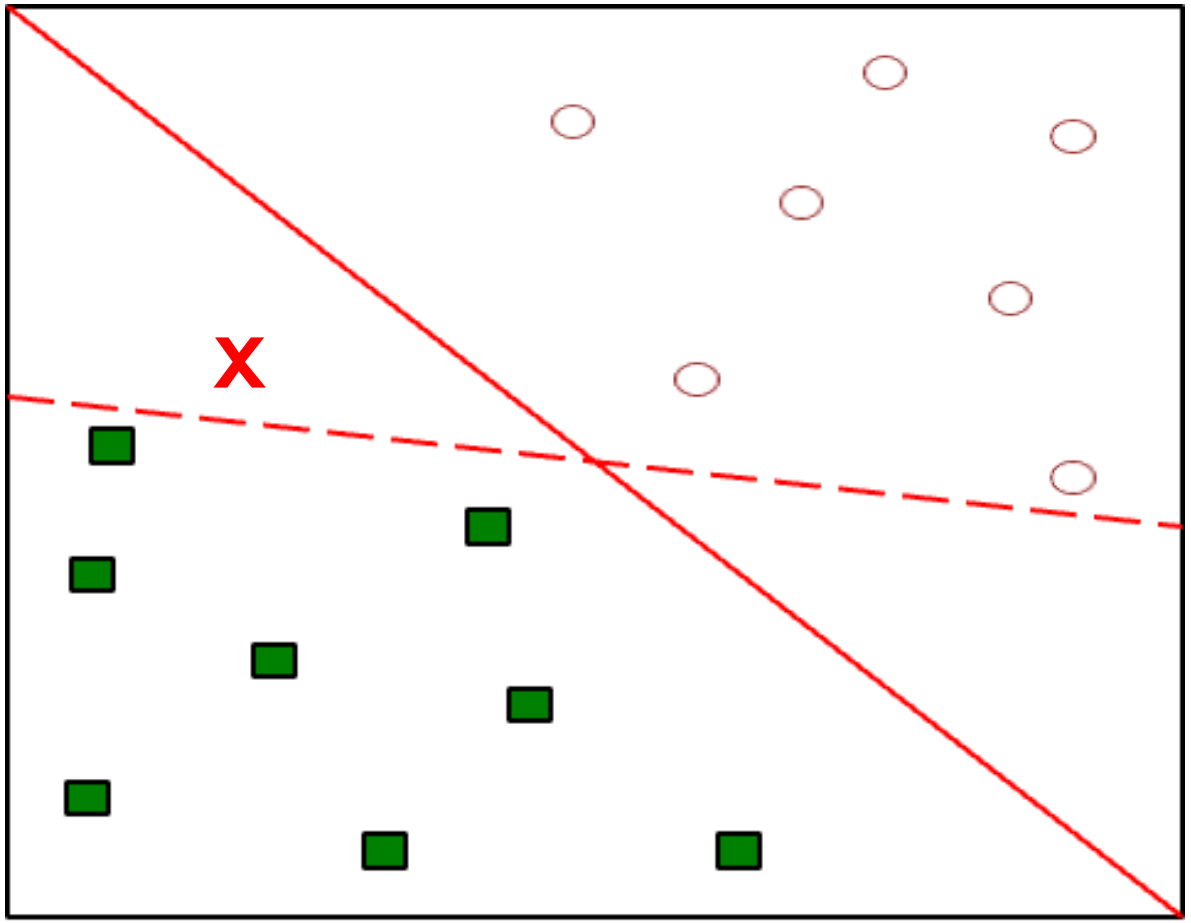
What's the best decision boundary?

How would you classify the new point?



What's the best decision boundary?

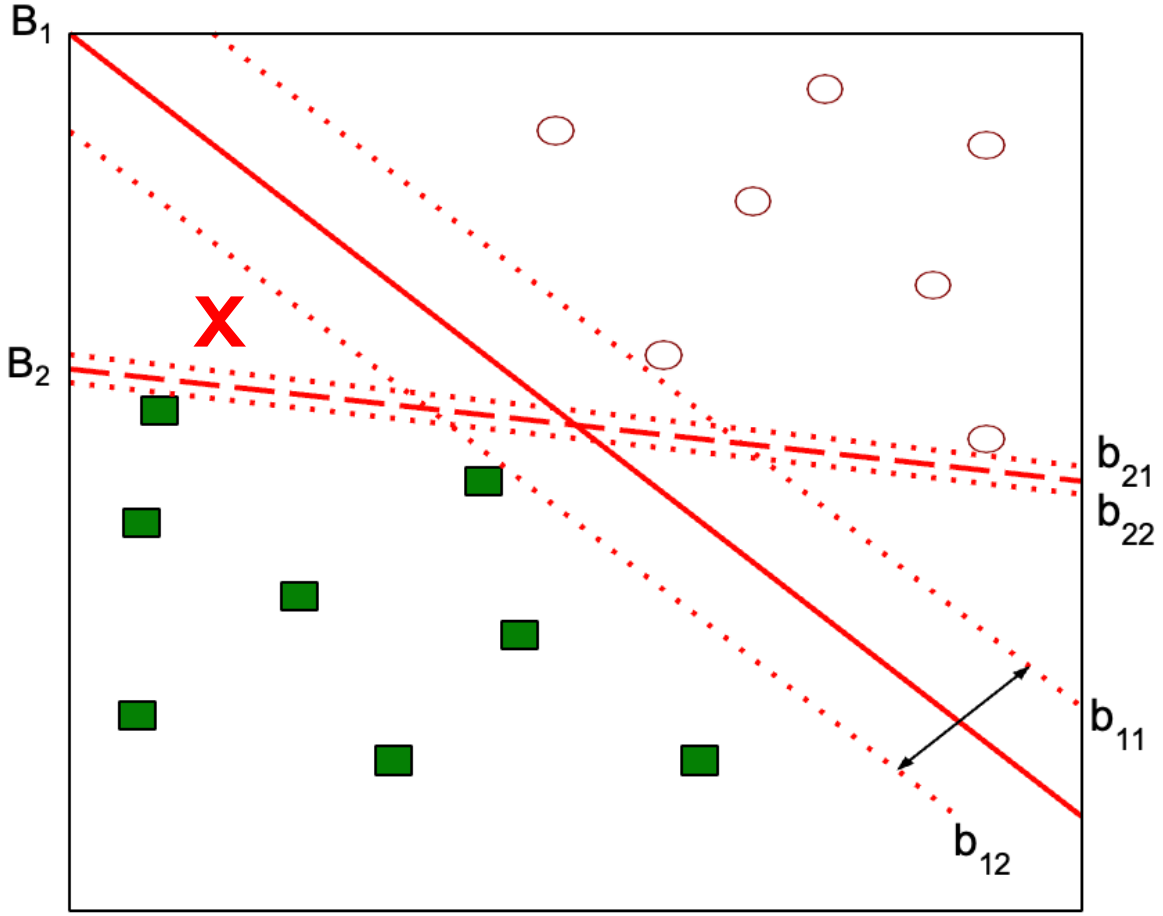
Which of the two hyperplanes matches your intuition?



What's the best decision boundary?

The best separating line (or hyperplane if there are more than 2 dimensions), is the one that **maximizes the margin.**

Classifiers that besides separating data points belonging to different classes also try to maximize the minimum distance of data points to the decision boundary are called **maximal margin classifiers.**



Support Vector Machines

SVMs

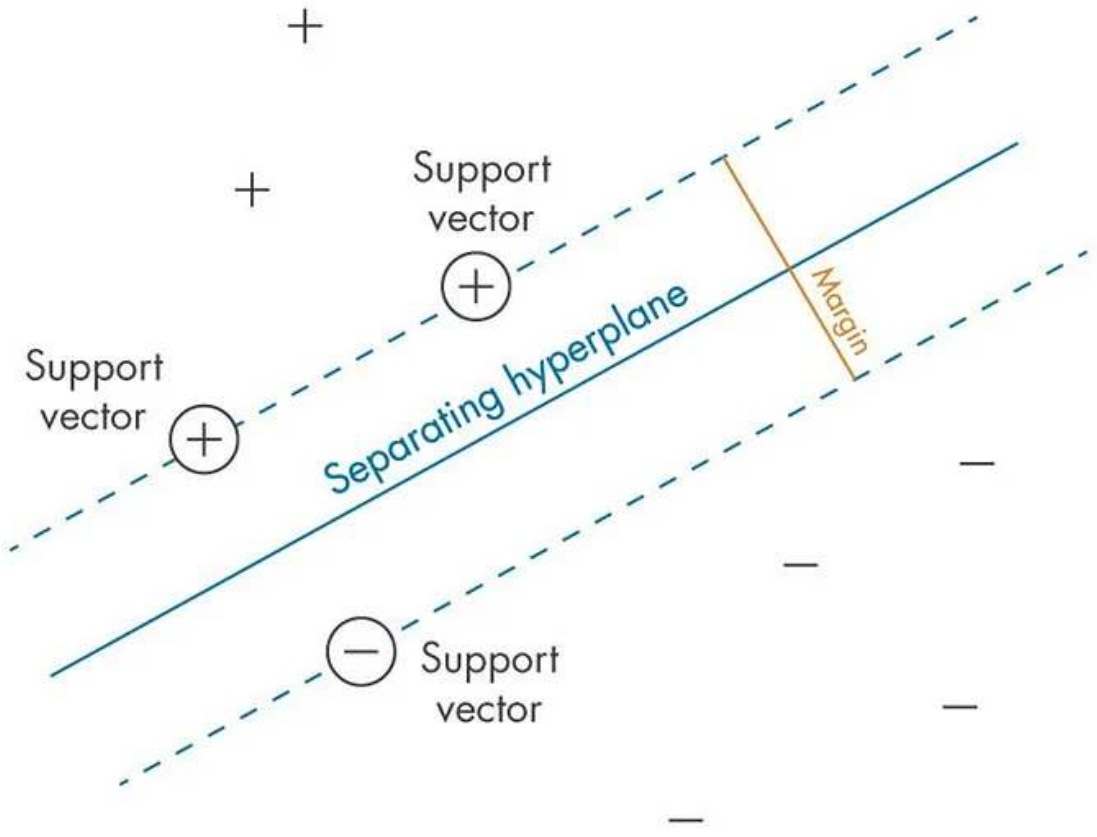
Support Vector Machines (SVMs) are supervised learning models used for classification (and regression).

They are particularly useful when:

- The number of features is large
- The number of samples is relatively small
- The boundary between classes is not obviously linear

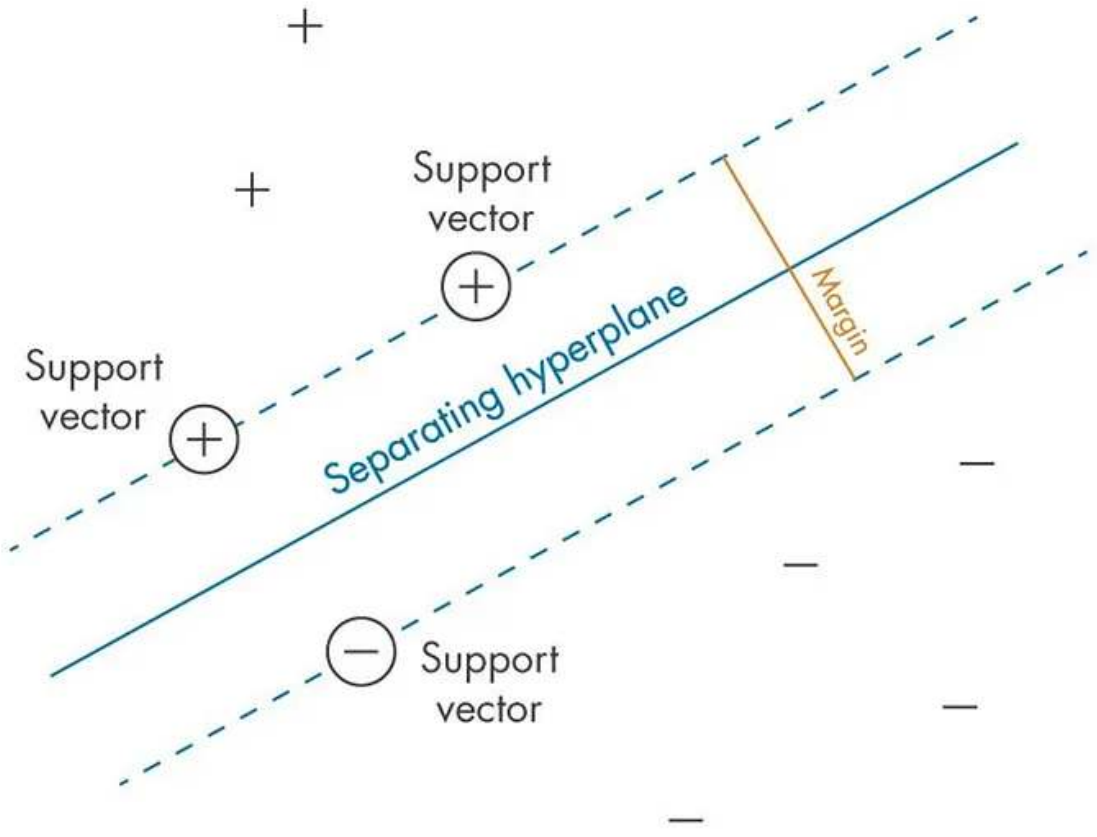
The Margin

- A margin is the distance between the decision boundary (hyperplane) and the closest data points from each class.
- The goal of SVMs is to maximize this margin.
- A larger margin indicates a greater degree of confidence in the classification.
- The margin is a measure of how well-separated the classes are.



Support Vectors

- Support vectors are the data points that lie closest to the decision boundary (hyperplane)
- They are important because they determine the position and orientation of the hyperplane.
- Support vectors are used to calculate the margin, which is the distance between the hyperplane and them.



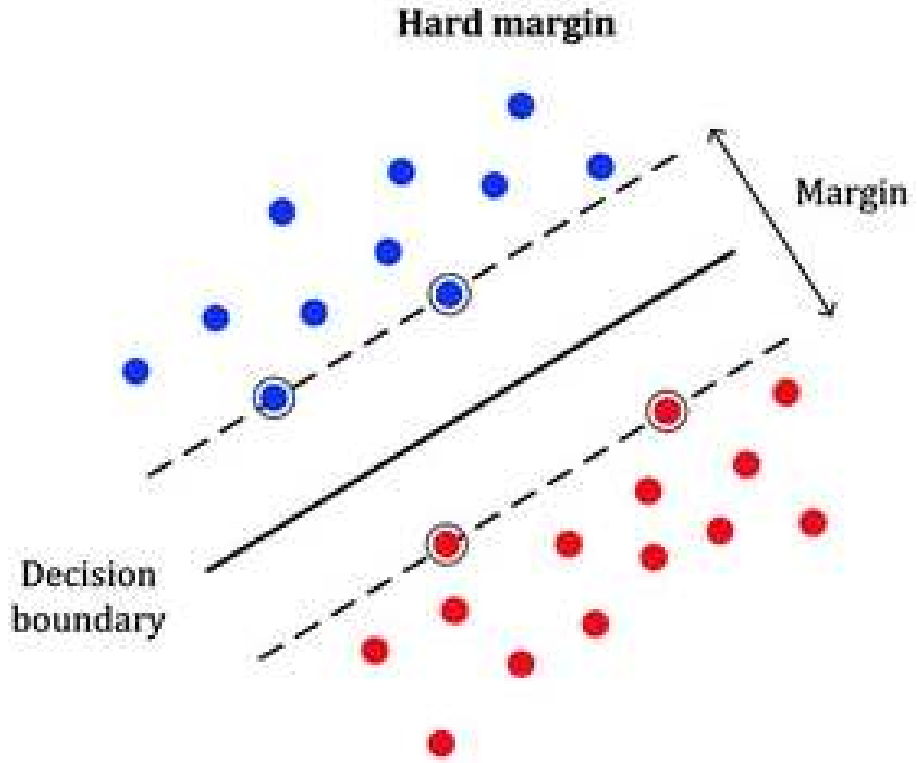
The Optimization Problem (Hard Margin)

To find the maximal margin hyperplane on data $(\mathbf{x}_i, y_i)$, with $y_i \in \{-1, 1\}$ solve:

$$\max_{w_0,\ldots,w_M} M$$

$$\text{subject to} \quad \sum_{j=0}^{M} w_j^2 = 1$$

$$y_i \left(\sum_{j=0}^{M} w_j x_{ij} \right) \geq M \quad \text{for all } i$$



This is a **constrained optimization problem** that can be solved using the Lagrange multiplier method.

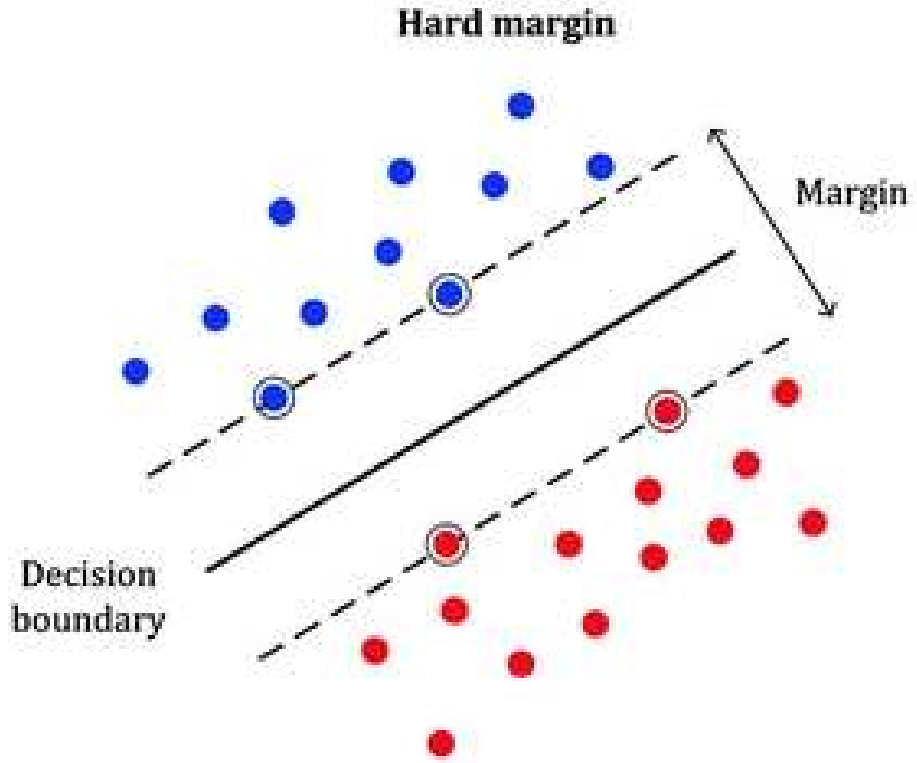
The Optimization Problem (Hard Margin)

To find the maximal margin hyperplane on data $(\mathbf{x}_i, y_i)$, with $y_i \in \{-1, 1\}$ solve:

$$\max_{w_0,\ldots,w_M} M \quad \leftarrow \text{Objective}$$

$$\text{subject to} \quad \sum_{j=0}^{M} w_j^2 = 1$$

$$y_i \left(\sum_{j=0}^{M} w_j x_{ij} \right) \geq M \quad \text{for all } i$$



This is a **constrained optimization problem** that can be solved using the Lagrange multiplier method.

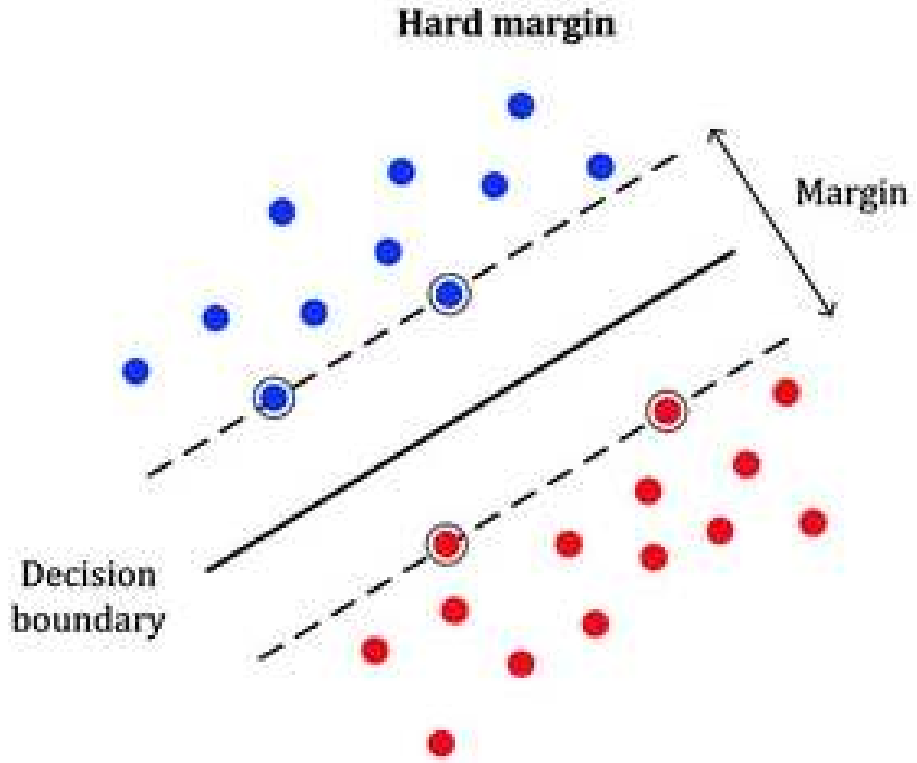
The Optimization Problem (Hard Margin)

To find the maximal margin hyperplane on data $(\mathbf{x}_i, y_i)$, with $y_i \in \{-1, 1\}$ solve:

$$\max_{w_0,\ldots,w_M} \quad M$$

$$\text{subject to} \quad \sum_{j=0}^{M} w_j^2 = 1 \quad \leftarrow \text{Fix Scale}$$

$$y_i \left(\sum_{j=0}^{M} w_j x_{ij} \right) \geq M \quad \text{for all } i$$



This is a **constrained optimization problem** that can be solved using the Lagrange multiplier method.

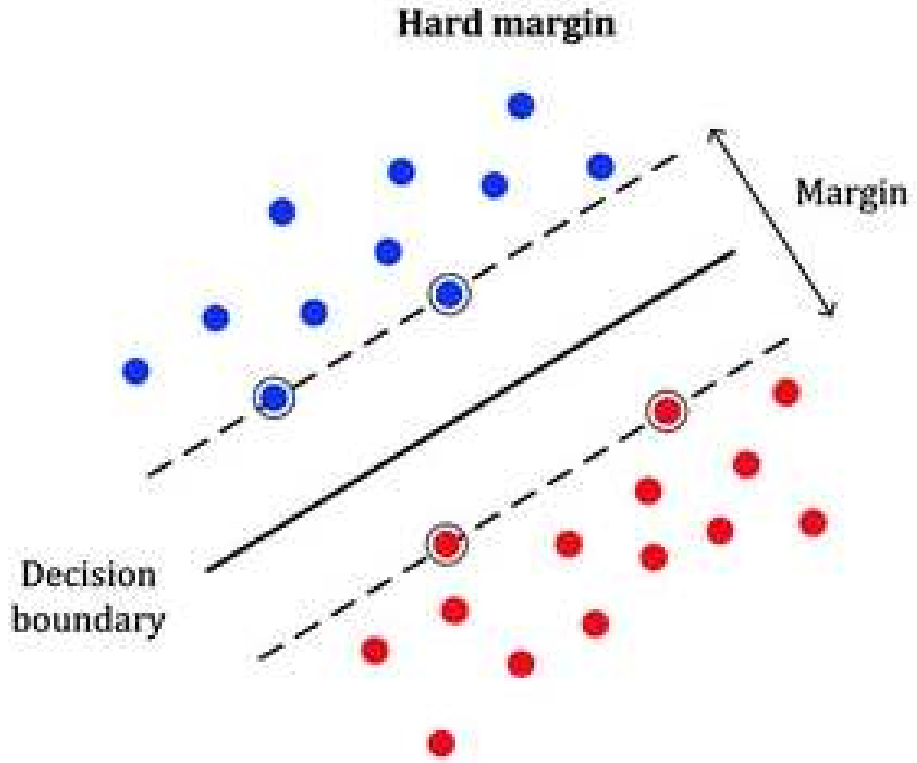
The Optimization Problem (Hard Margin)

To find the maximal margin hyperplane on data $(\mathbf{x}_i, y_i)$, with $y_i \in \{-1, 1\}$ solve:

$$\max_{w_0,\ldots,w_M} \quad M$$

$$\text{subject to} \quad \sum_{j=0}^{M} w_j^2 = 1$$

$$\underbrace{y_i \left(\sum_{j=0}^{M} w_j x_{ij} \right)}_{\text{functional margin}} \geq M \quad \text{for all } i$$



This is a **constrained optimization problem** that can be solved using the Lagrange multiplier method.

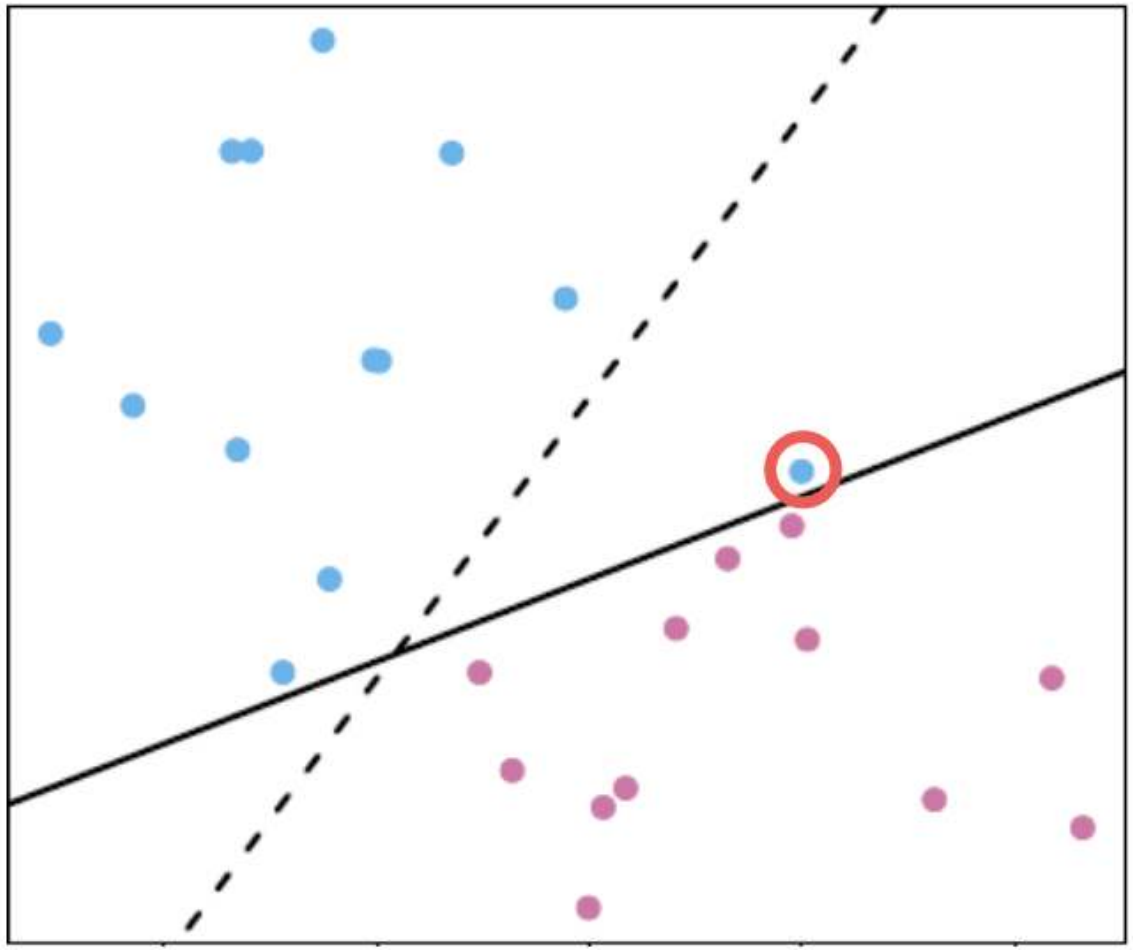
The Problem with Hard Margins

Hard margins have a few problems:

- Can be sensitive to individual observations (e.g. outliers)
- May overfit training data

We could use instead a soft margin that:

- allows training samples to be on the "wrong side" of the margin or hyperplane.
- defines a hyperplane that **almost** separates the classes.



The Optimization Problem (Soft Margin)

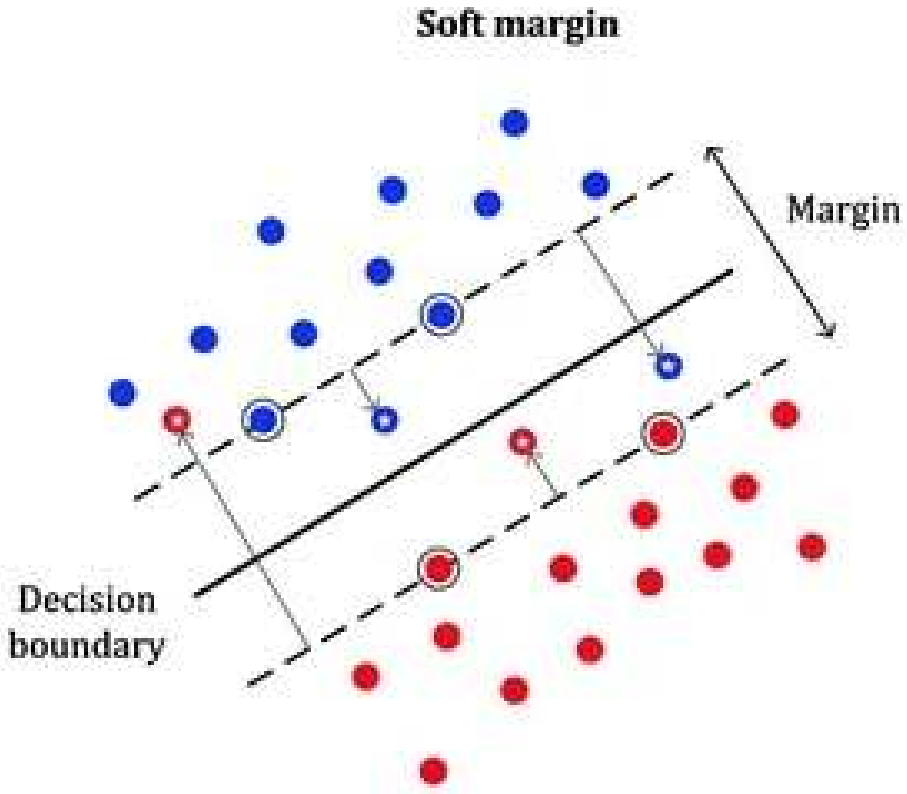
To allow points on the wrong side of the margin we introduce a slack variable ϵ_i:

$$\max_{w_0,\ldots,w_M,\epsilon_1,\ldots,\epsilon_N} M$$

$$\text{subject to} \quad \sum_{j=0}^{M} w_j^2 = 1$$

$$y_i \left(\sum_{j=0}^{M} w_j x_{ij} \right) \geq M(1 - \epsilon_i)$$

$$\underbrace{\sum_{i=1}^{N} \epsilon_i \leq C}_{\text{Slack "budget"}} \quad \epsilon_i \geq 0 \quad \text{for all } i$$



- Large C: allow more violations. Small C: allow fewer violations.

The Kernel Trick

What if the decision boundary is not linear?

- The original feature space can always be mapped to some **higher-dimensional** feature space where the training set is separable.
- The trick is to **transform the data** from its original space $\mathbf{x}$ into a new space $\phi(\mathbf{x})$ ($\mathbf{x} \mapsto \phi(\mathbf{x})$) so that a linear decision boundary can be used.

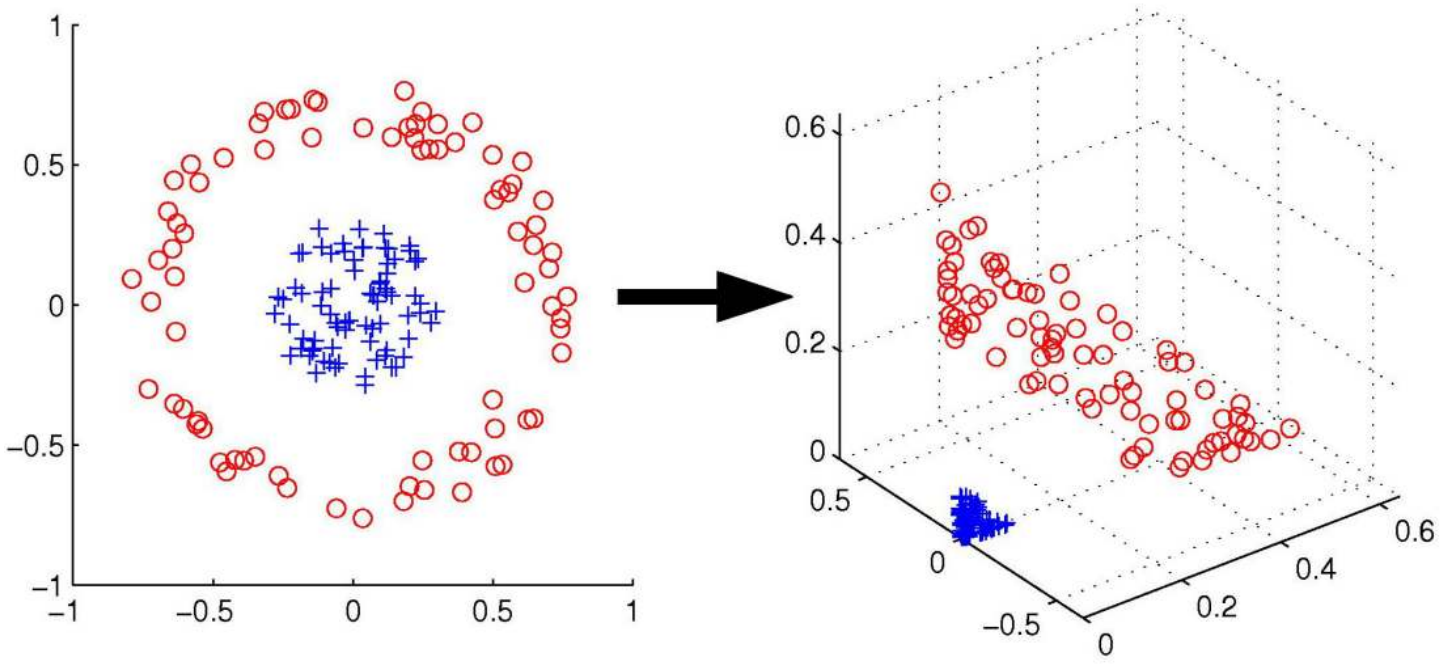


Figure 1: Transforming the data can make it linearly separable

The Kernel Trick

In some cases, explicitly computing $\phi(x)$ would be computationally infeasible. Instead of computing $\phi(\mathbf{x})$ explicitly, use a kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j)$$

It behaves as if we had mapped data to a higher dimensional space, without ever computing that space explicitly. This allows for finding efficient nonlinear decision boundaries.

Common kernels are **Linear**, **Polynomial**, Radial Basis Function (**RBF**)

Summary and Limitations

- SVM finds a separating boundary with **maximum margin**
- Only a subset of points (support vectors) determine the solution
- Soft margin allows controlled violations via slack variables
- Kernels allow nonlinear decision boundaries
- Works well in high-dimensional settings such as genomics

Limitations

- Choice of kernel and hyperparameters (e.g., C) is critical
- Less interpretable than simple linear models
- Training can be slow for very large datasets

In many biological applications linear SVM is often surprisingly competitive, but careful cross-validation is essential, and feature scaling is mandatory.

SVMs can also be used for regression!

Exercise

Given the following set of data, roughly sketch the optimal separating hyperplane, and, based on your choice, identify the support vectors.

x_1	x_2	y
1	0	+1
3	7	+1
6	10	+1
3	2	-1
5	6	-1
8	3	-1