SUPPORT VECTOR MACHINE INTRODUCTION

Pei-Yuan Wu
Electrical Engineering Department
National Taiwan University

OUT-LINE

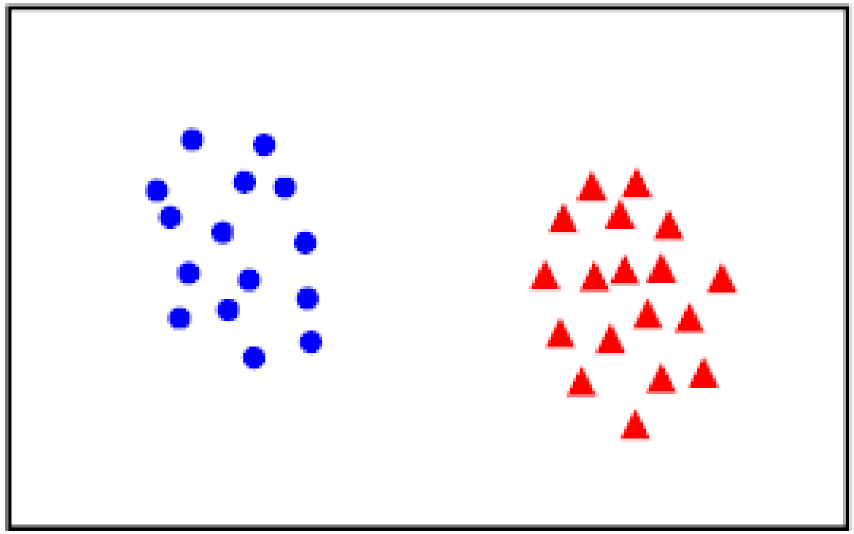
- Review of linear classifiers
  - Linear separability
- Support Vector Machine (SVM) classifier
  - The role of margin
  - Optimal margin hyperplanes
  - Soft Margin SVM and Slack variables

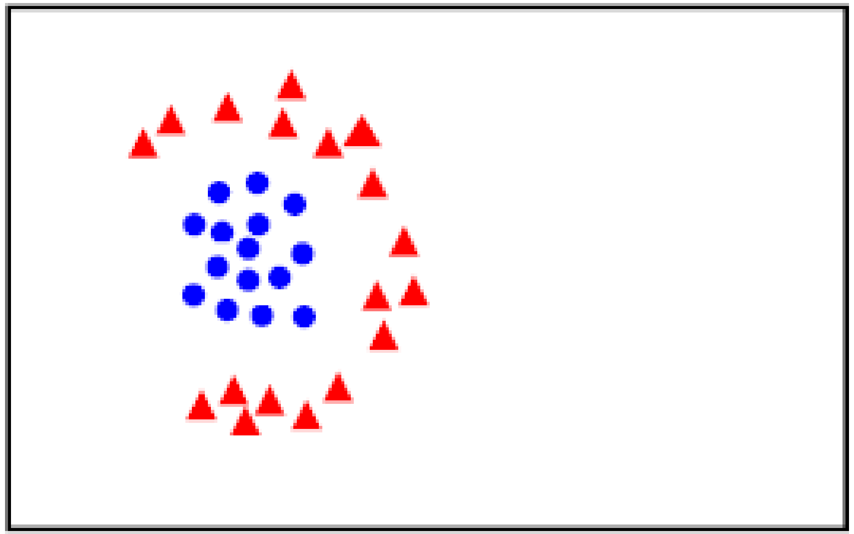
BINARY CLASSIFICATION

- Given training data $(\boldsymbol{x}_i, y_i)$ for $i = 1...N$, with $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$, learn a classifier f(x) such that

$$f(\boldsymbol{x}_i) \begin{cases} \geq 0 & , if\ y_i = +1 \\ < 0 & , if\ y_i = -1 \end{cases}$$

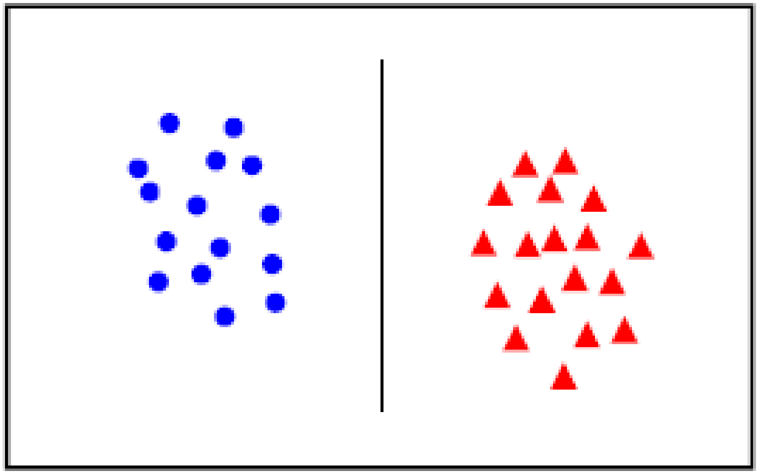
i.e. $y_i f(\boldsymbol{x}_i) > 0$ for a correct classification.

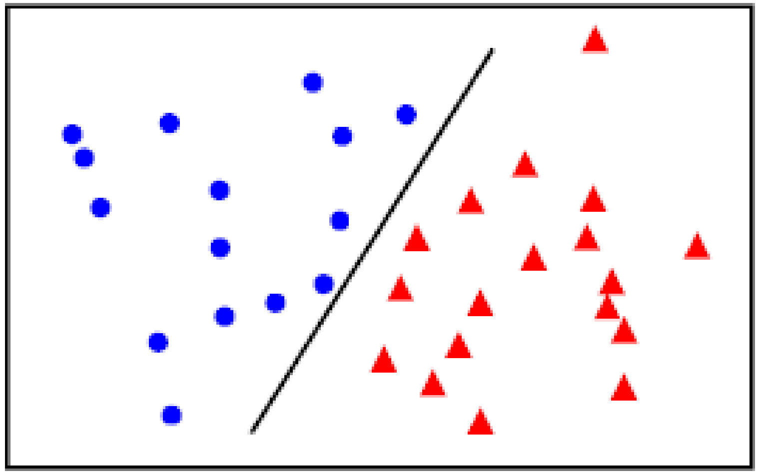




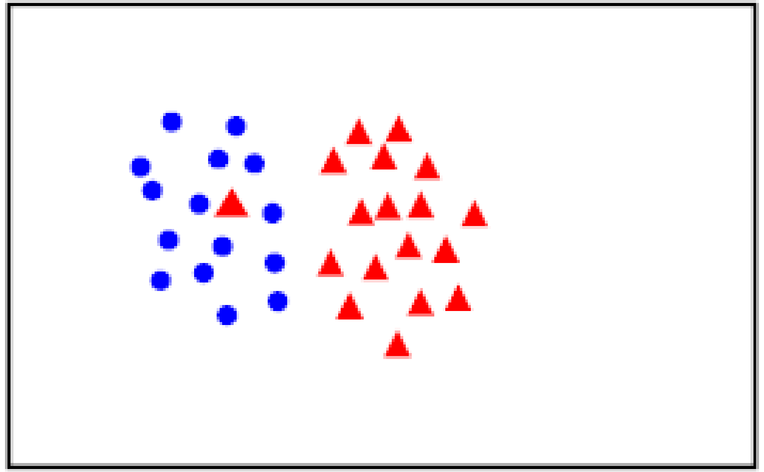
LINEAR SEPARABILITY

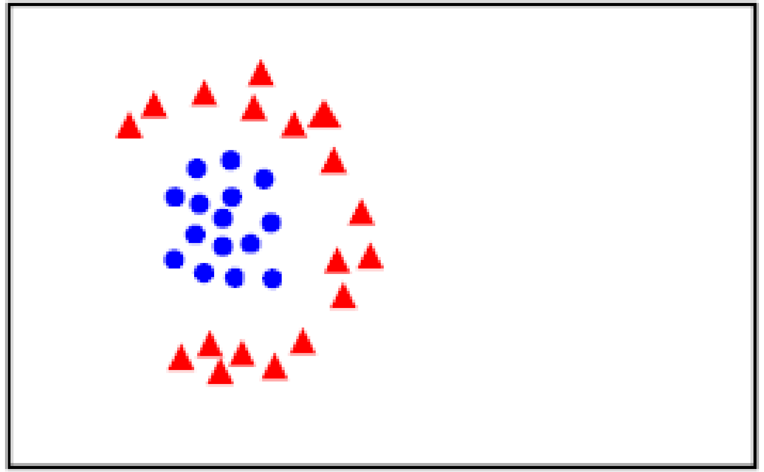
Linearly separable





Not
Linearly separable



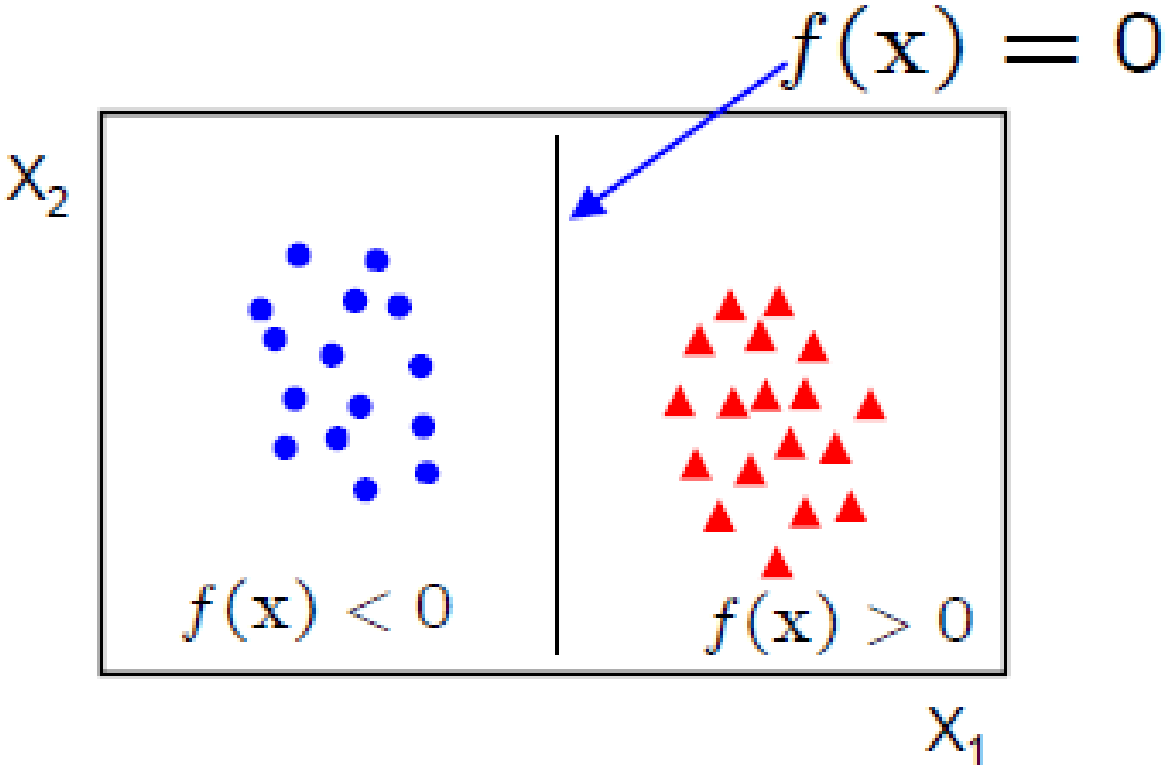


LINEAR CLASSIFIERS

- A linear classifier has the form

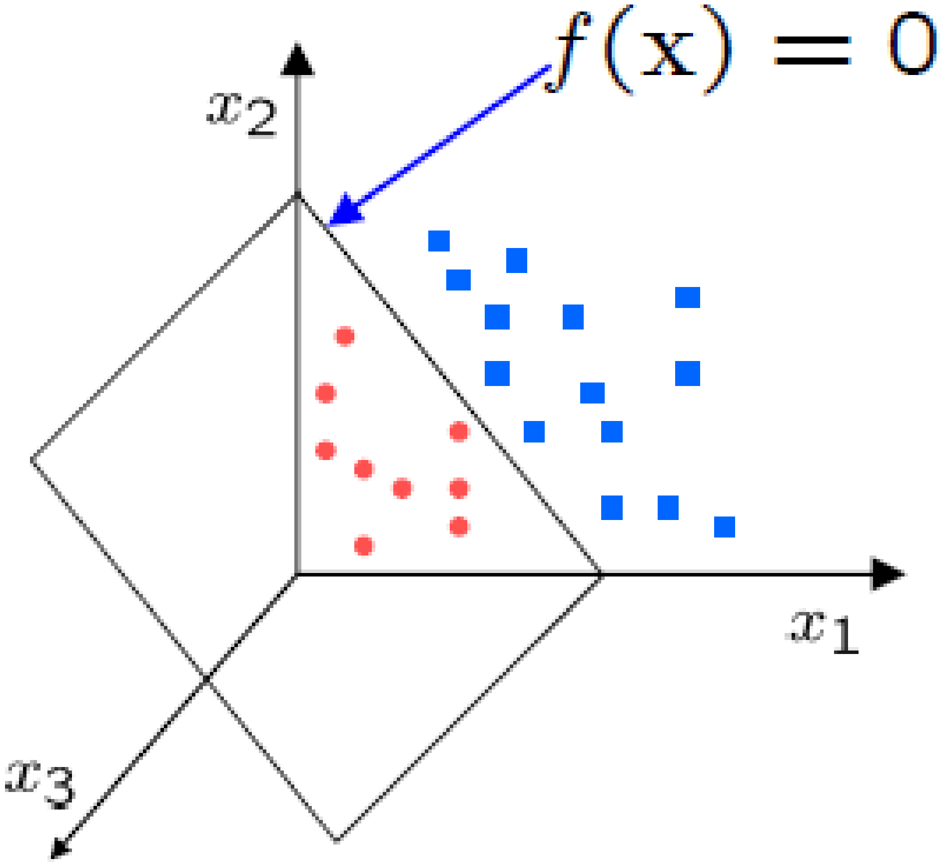
$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b$$

- in 2D the discriminant is a line
- w is the normal to the line, and b the bias
- w is known as the weight vector



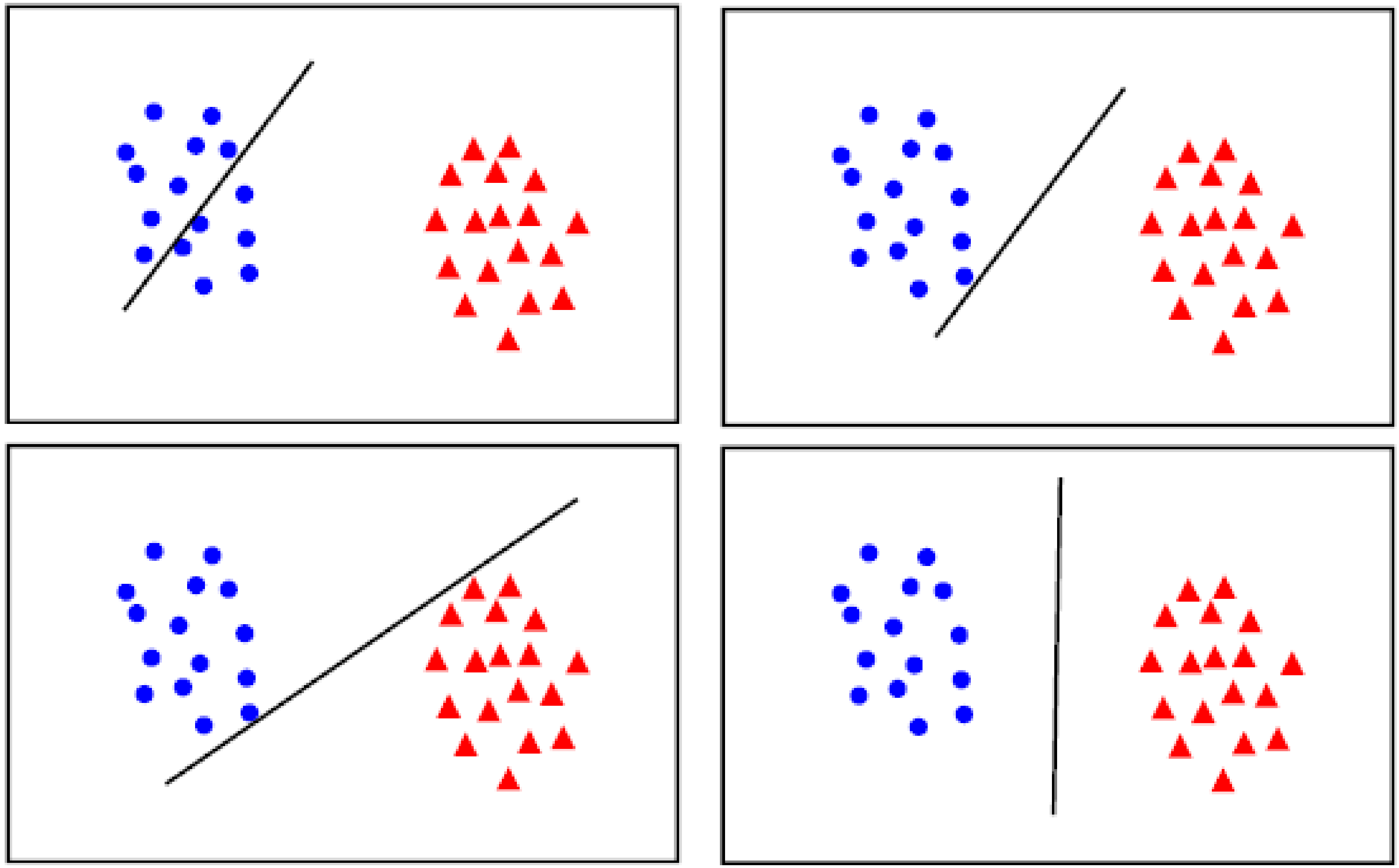
LINEAR CLASSIFIERS

- A linear classifier has the form
$$f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b$$
- in 3D the discriminant is a plane, and in n-D it is a hyperplane
- For a K-NN classifier it was necessary to `carry' the training data
- For a linear classifier, the training data is used to learn **w** and then discarded
- Only **w** is needed for classifying new data



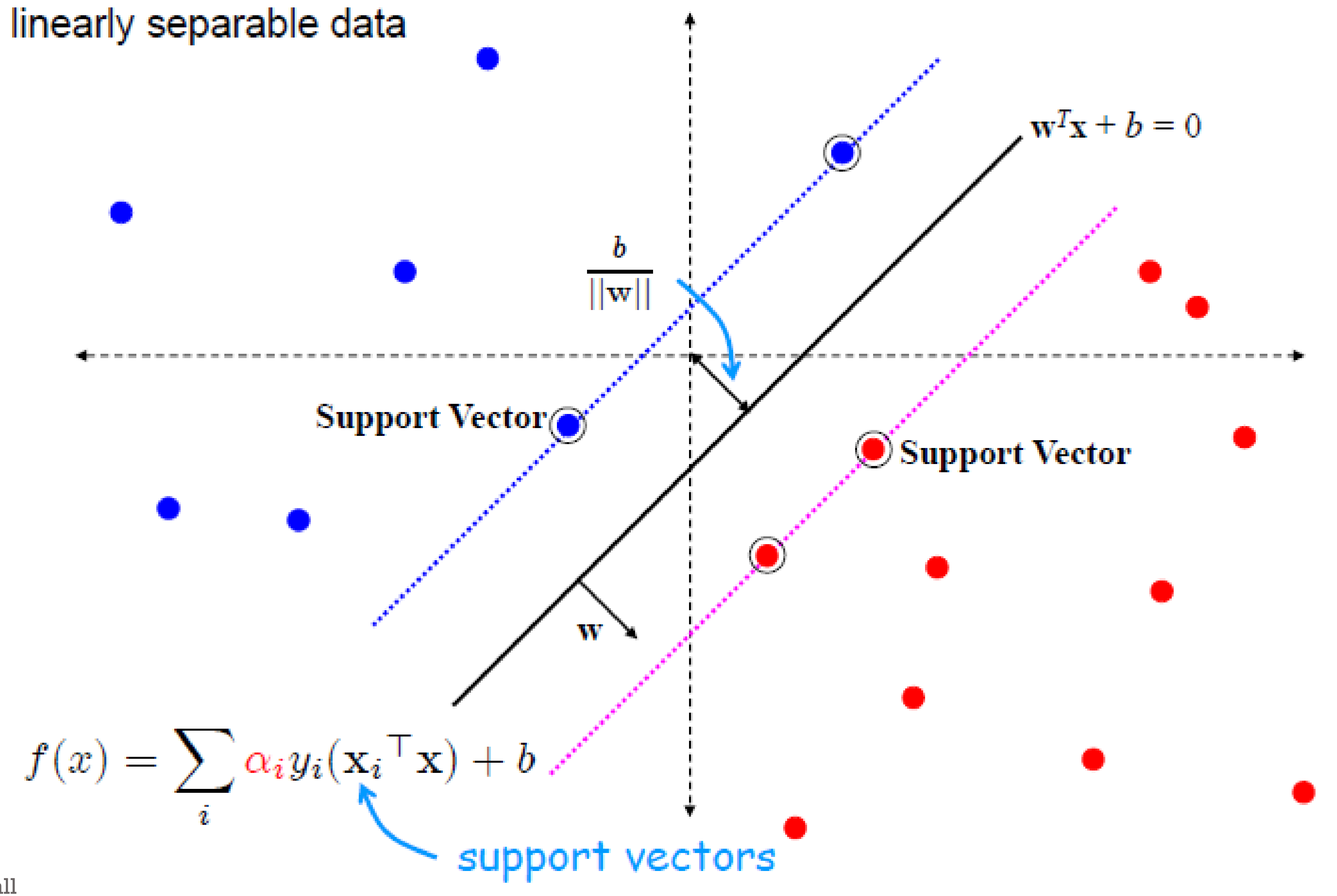
THE ROLE OF MARGIN

What is the best w?



maximum margin solution: most stable under perturbations of the inputs

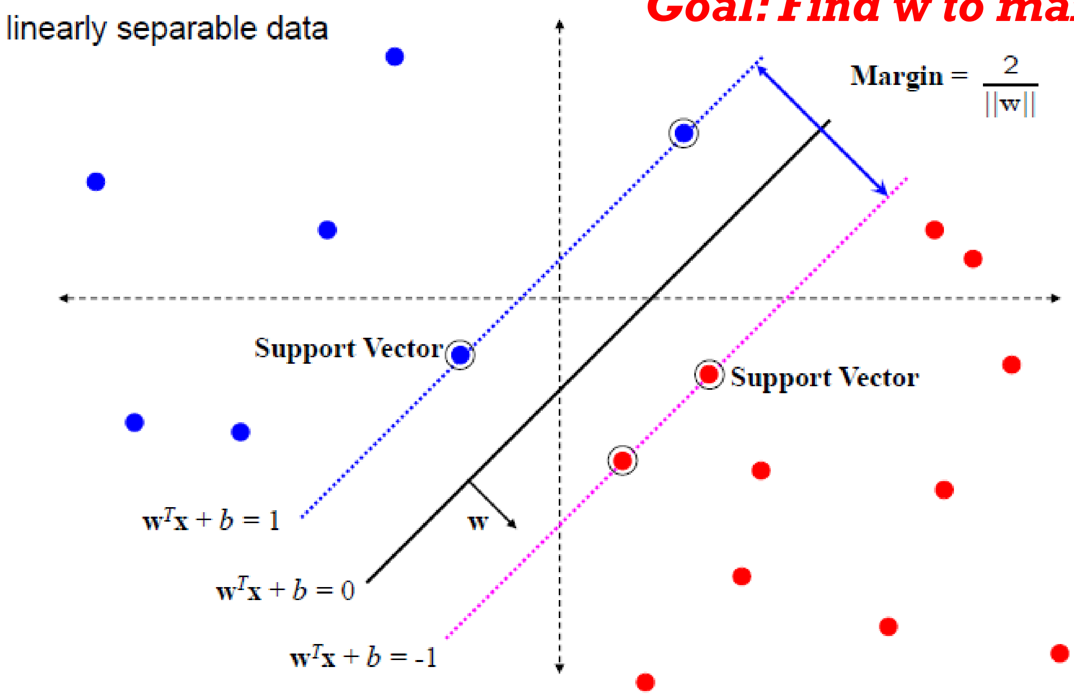
SUPPORT VECTOR MACHINE



OPTIMAL MARGIN HYPERPLANES

- Since $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ and $c(\boldsymbol{w}^T\boldsymbol{x} + b) = 0$ define the same plane, we have the freedom to choose the normalization of $(\boldsymbol{w}, b)$
- Choose normalization such that
  - $\boldsymbol{w}^T\boldsymbol{x}_+ + b = +1$ for the positive support vectors $\boldsymbol{x}_+$
  - $\boldsymbol{w}^T\boldsymbol{x}_- + b = -1$ for the negative support vectors $\boldsymbol{x}_-$
- Then the margin is given by $\frac{2}{\|\boldsymbol{w}\|}$

Goal: Find w to maximize margin



SVM – OPTIMIZATION

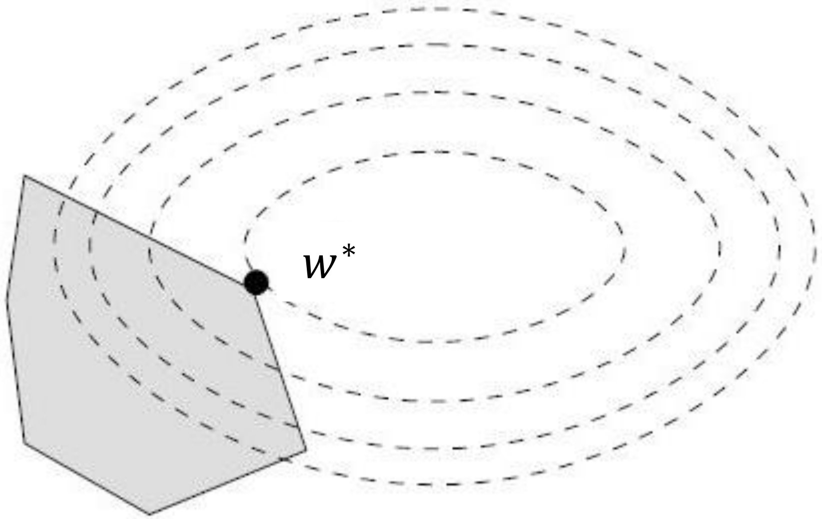
- Learning the SVM can be formulated as an optimization:

$$\text{Maximize } \frac{2}{\|\boldsymbol{w}\|} \text{ subject to } \boldsymbol{w}^T\boldsymbol{x}_i + b \begin{cases} \geq 1, if\ y_i = +1 \\ \leq -1, if\ y_i = -1 \end{cases}, \text{for } i = 1, \dots, n$$

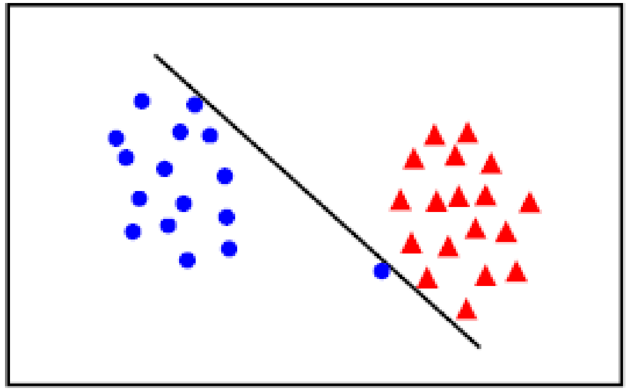
- Or equivalently

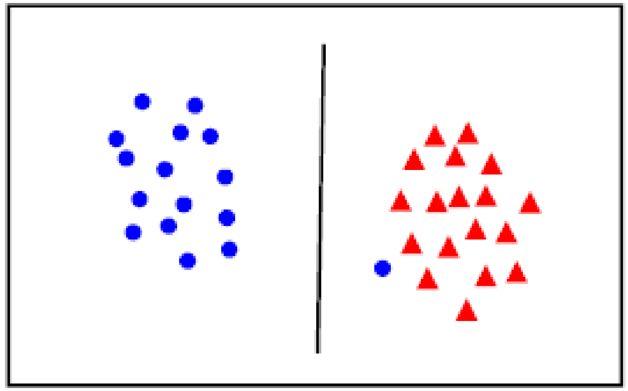
$$\text{Minimize } \frac{1}{2}\|\boldsymbol{w}\|^2 \text{ subject to } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, \text{for } i = 1, \dots, n$$

- This is a quadratic optimization problem subject to linear constraints and there is a unique minimum



LINEAR SEPARABILITY AGAIN: WHAT IS THE BEST W?





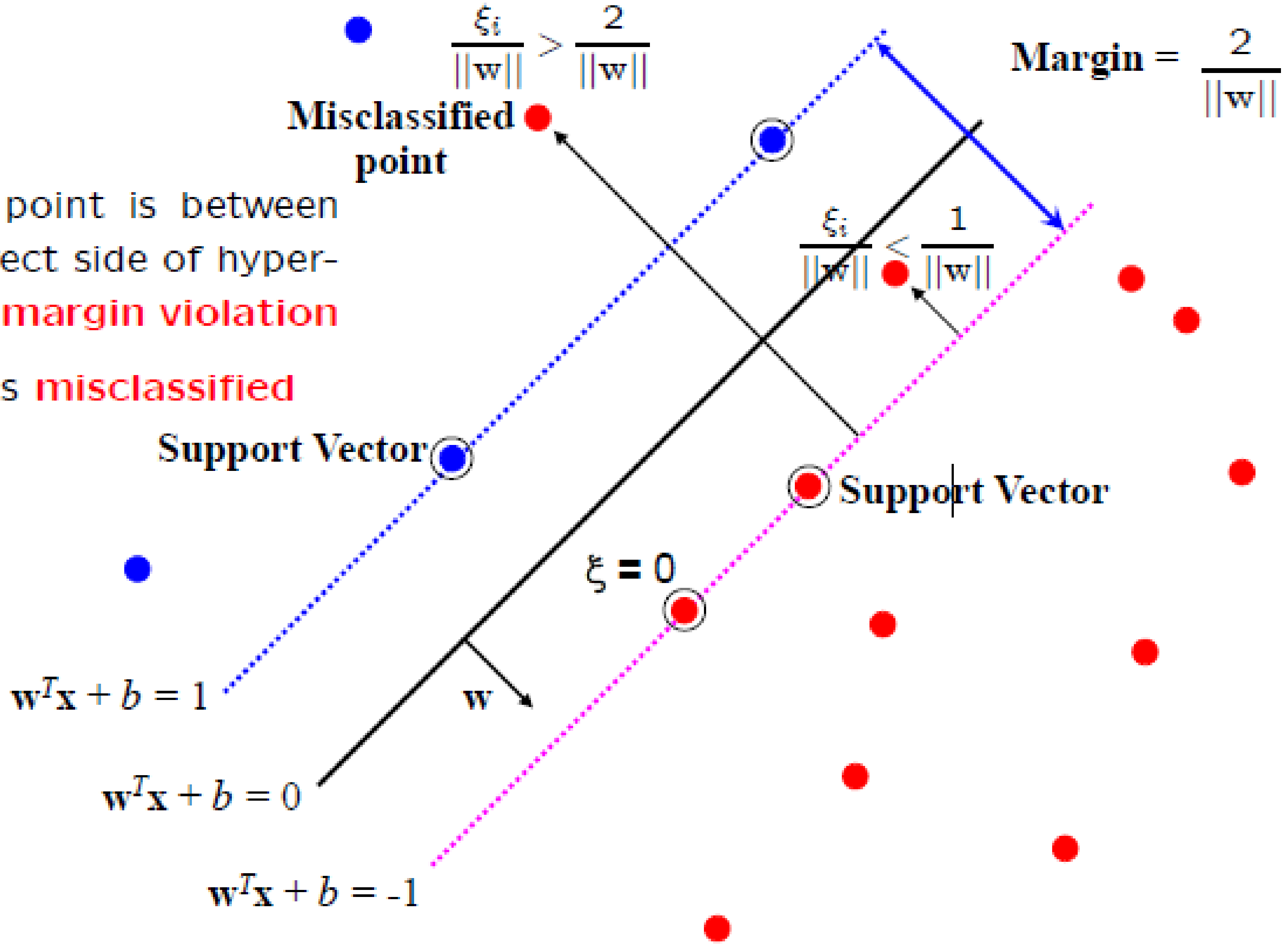
- the points can be linearly separated but there is a very narrow margin
- but possibly the large margin solution is better, even though one constraint is violated

In general there is a trade off between the margin and the number of mistakes on the training data

INTRODUCE "SLACK" VARIABLES

$$\xi_i \geq 0$$

- for $0 < \xi \leq 1$ point is between margin and correct side of hyper-plane. This is a margin violation
- for $\xi > 1$ point is misclassified



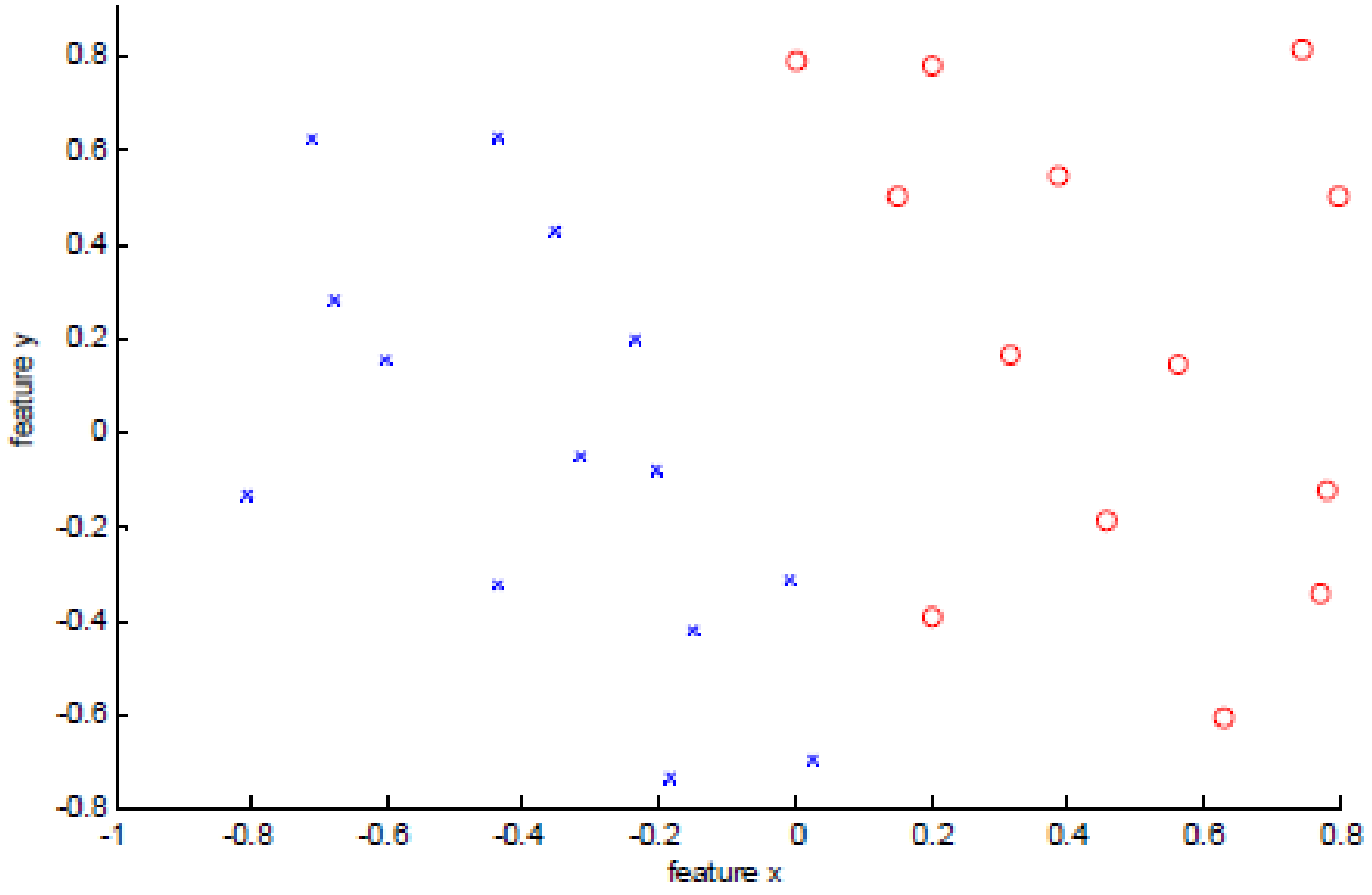
"SOFT" MARGIN SVM

- The optimization problem becomes

$$\min_{\boldsymbol{w}\in\mathbb{R}^{\boldsymbol{d}}, b\in\mathbb{R}, \xi_1,\dots,\xi_N\geq 0} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

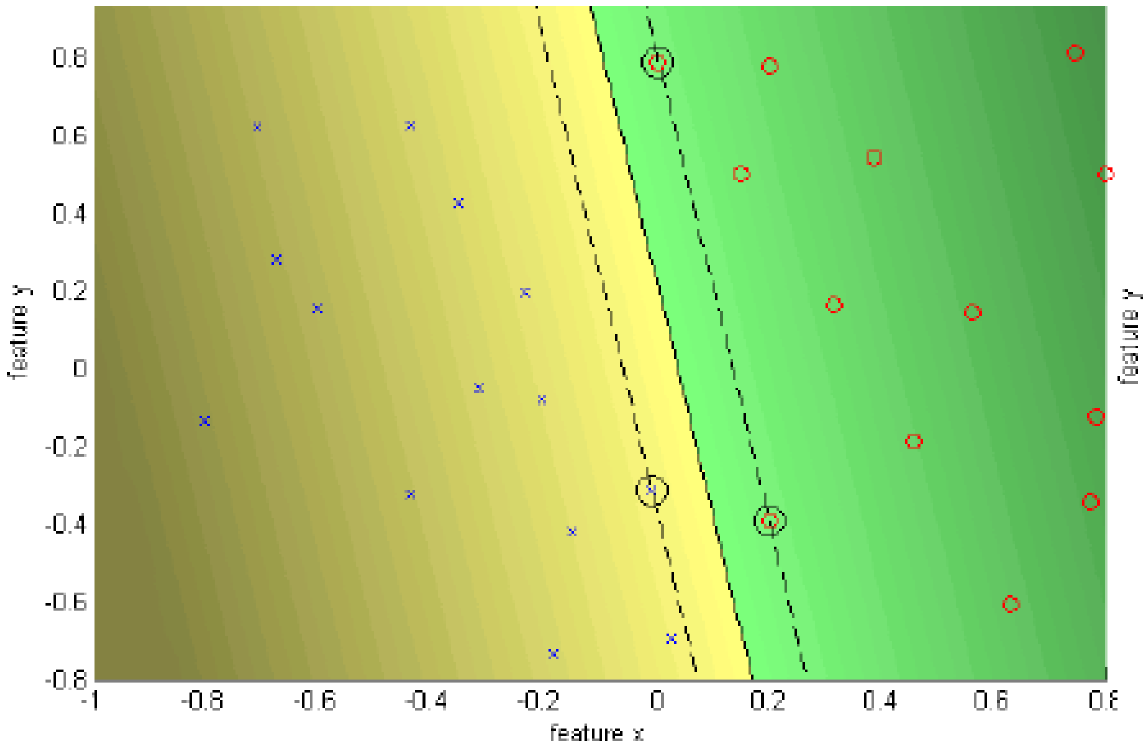
subject to $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i$, for $i = 1, \dots, n$

- Every constraint can be satisfied if ξ_i is sufficiently large
- C is a regularization parameter:
  - small C allows constraints to be easily ignored → large margin
  - large C makes constraints hard to ignore → narrow margin
  - $C = \infty$ enforces $\xi_i = 0$ for $i = 1, \dots, n$ → hard margin
- This is still a quadratic optimization problem and there is a unique minimum. Note, there is only one parameter, C.



- data is linearly separable
- but only with a narrow margin

C = Infinity → hard margin



C = 10 → soft margin