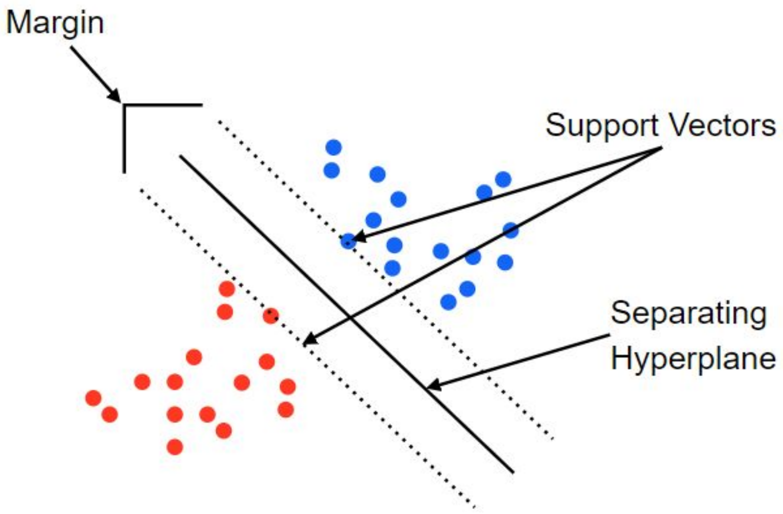
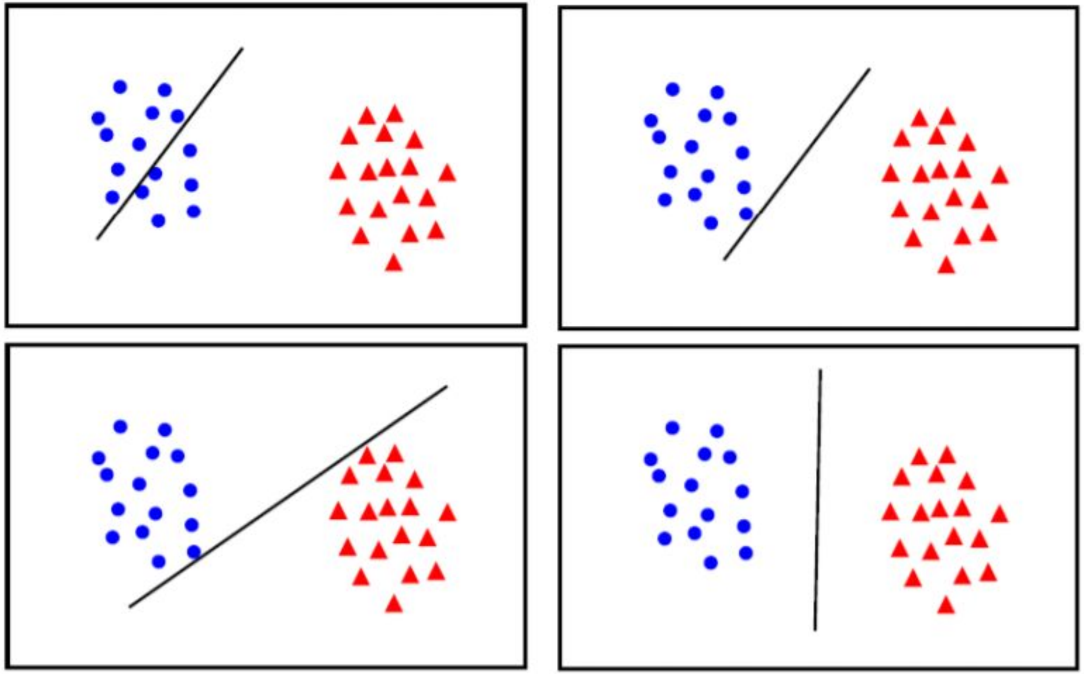
Chapter 3.4 SVM

What is the support vector machine

1. Task: Two class classification
2. Goal: Find optimal hyperplane to separate two groups
    a. The optimum state is when the distances of the SVs of the two classes become maximum.



Why maximize margin?



- Maximum margin solution: most stable under perturbations of the inputs

Decision rule

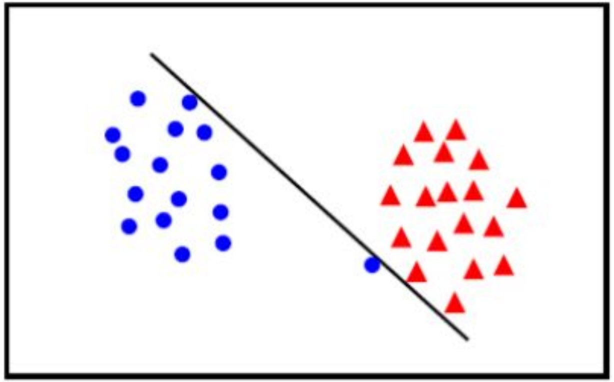
- Learning the SVM can be formulated as an optimization:

$$\max_{w} \frac{2}{\|w\|} \text{ subject to } w^T x_i + b \begin{cases} \geq \ \ 1 & if\ y_i = +1 \\ \leq -1 & if\ y_i = -1 \end{cases} \text{ for } i = 1 \dots N$$

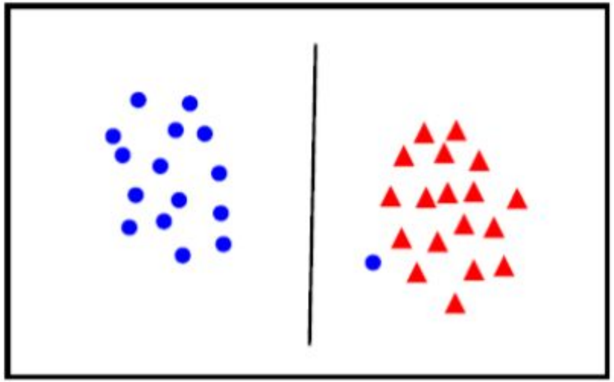
- Or equivalently

$$\min_{w} \|w\|^2 \text{ subject to } y_i(w^T x_i + b) \geq 1 \text{ for } i = 1 \dots N$$

What is the best normal vector 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

Soft margin classification

Slack variable

The optimization problem becomes

$$\min_{w \in Rd, \xi_i \in R^+} \|w\|^2 + \boxed{c \sum_{i}^{N} \xi}$$

subject to

$$y_i(w^T x_i + b) \geq 1 - \xi_i \text{ for } i = 1 \ldots N$$

Optimization

Learning an SVM has been formulated as a constrained optimization problem over $\mathbf{w}$ and ξ

$$\min_{\mathbf{w}\in\mathbb{R}^d, \xi_i \in \mathbb{R}^+} ||\mathbf{w}||^2 + C\sum_i^N \xi_i \text{ subject to } y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i \text{ for } i = 1 \ldots N$$

The constraint $y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i$, can be written more concisely as

$$y_i f(\mathbf{x}_i) \geq 1 - \xi_i$$

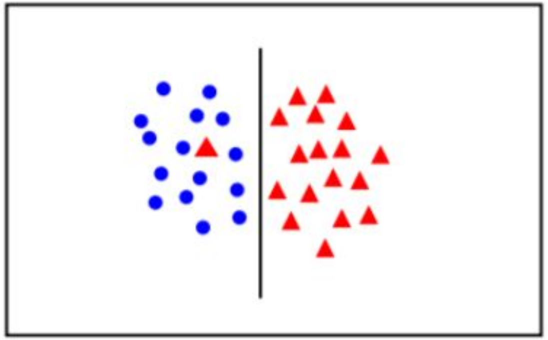
which, together with $\xi_i \geq 0$, is equivalent to

$$\xi_i = \max\left(0, 1 - y_i f(\mathbf{x}_i)\right)$$

Hence the learning problem is equivalent to the unconstrained optimization problem over $\mathbf{w}$

$$\min_{\mathbf{w}\in\mathbb{R}^d} \underbrace{||\mathbf{w}||^2}_{\text{regularization}} + C\sum_i^N \underbrace{\max\left(0, 1 - y_i f(\mathbf{x}_i)\right)}_{\text{loss function}}$$

Non-separable problem

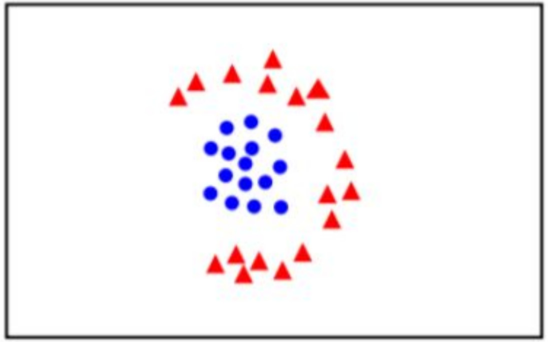


- introduce slack variables

$$\min_{\mathbf{w}\in\mathbb{R}^d,\xi_i\in\mathbb{R}^+} ||\mathbf{w}||^2 + C\sum_i^N \xi_i$$

subject to

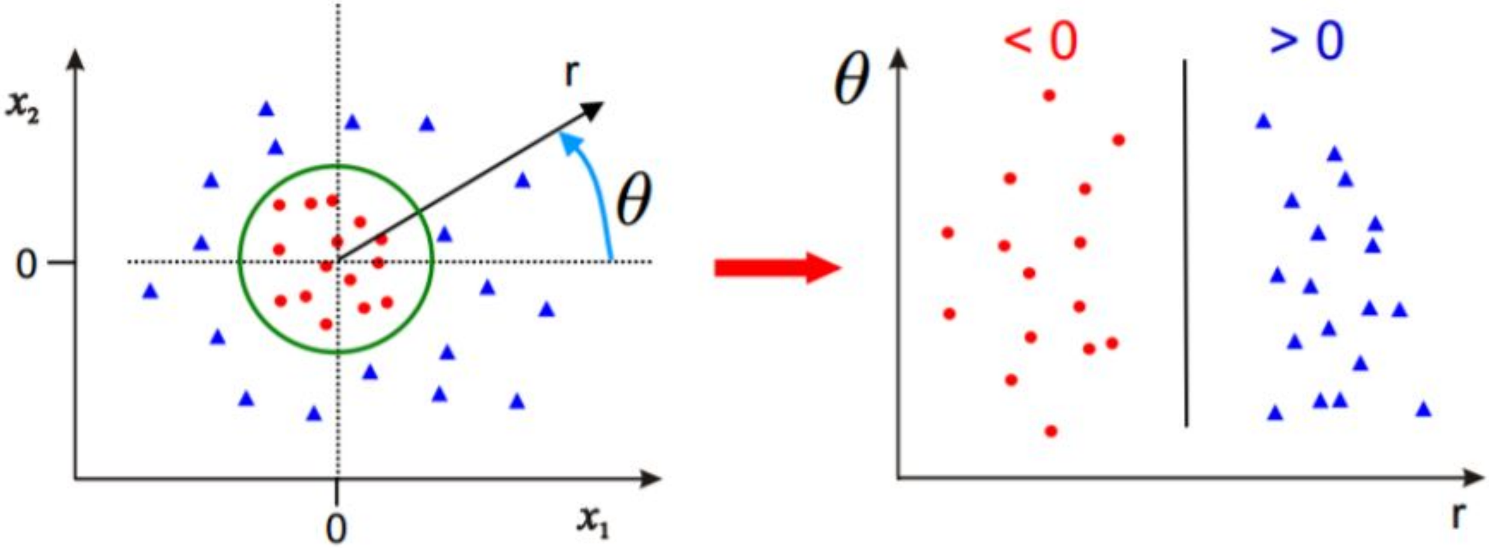
$$y_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i \text{ for } i = 1 \ldots N$$



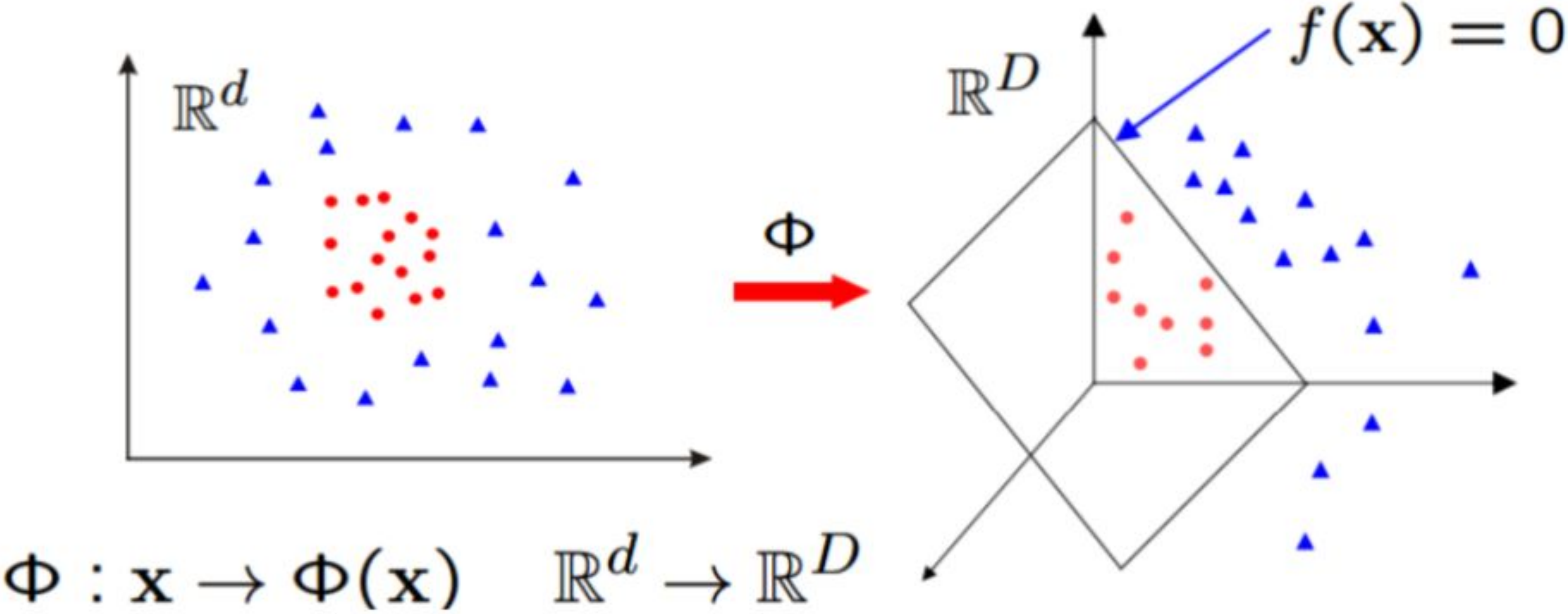
- linear classifier not appropriate

??

Solution 1: Polar coordinate transformation



Solution 2: Feature space transformation



Kernel trick

Linear kernels $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$

Polynomial kernels $k(\mathbf{x}, \mathbf{x}') = \left(1 + \mathbf{x}^\top \mathbf{x}'\right)^d$ for any $d > 0$

– Contains all polynomials terms up to degree d

Gaussian kernels $k(\mathbf{x}, \mathbf{x}') = \exp\left(-||\mathbf{x} - \mathbf{x}'||^2 / 2\sigma^2\right)$ for $\sigma > 0$

– Infinite dimensional feature space