NONLINEAR SVM

What if the training set is NOT LINEARLY SEPARABLE?

$C_1 = \{ x_1, \ldots, x_m \} \Rightarrow d_i = +1 \quad i = 1, \ldots, m$

$C_2 = \{ x_{m+1}, \ldots, x_N \} \Rightarrow d_i = -1 \quad i = m+1, \ldots, N$

Question: Let $\underset{\sim}{x} \in \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$. Can we find a TRANSFORMATION

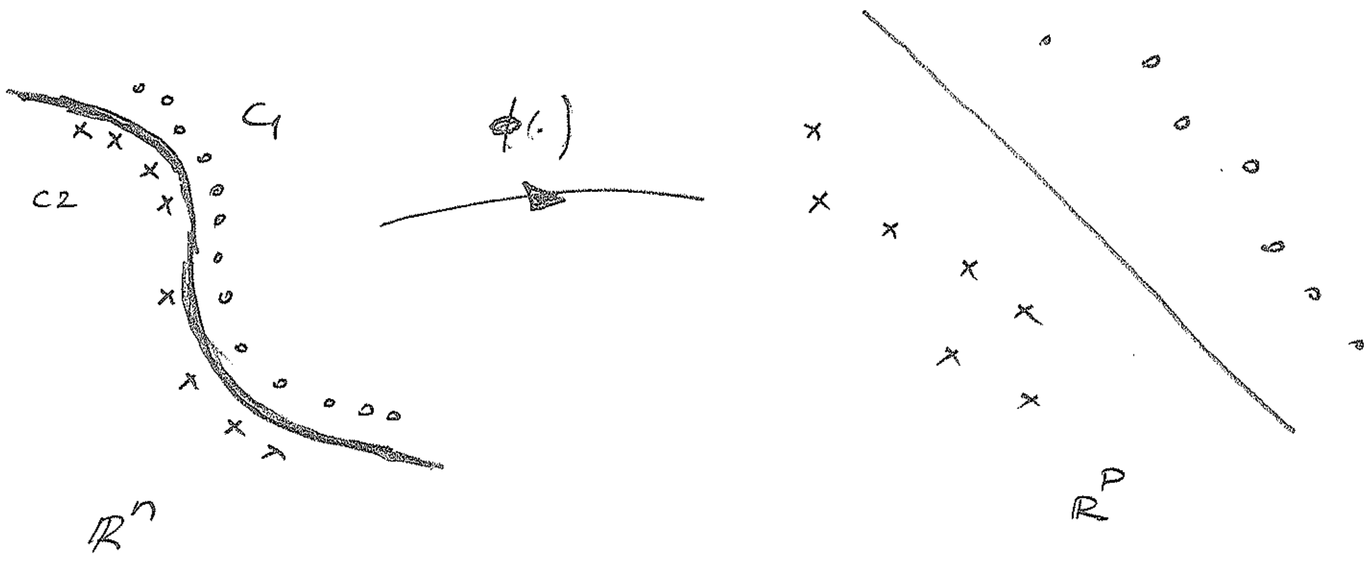
$$\phi(x) = \begin{pmatrix} \phi_1(x) \\ \vdots \\ \phi_p(x) \end{pmatrix} \in \mathbb{R}^p$$

such that

$$\phi(C_1) = \{ \phi(x_1), \ldots, \phi(x_m) \} \Rightarrow d_i = +1$$

$$\phi(C_2) = \{ \phi(x_{m+1}), \ldots, \phi(x_N) \} \Rightarrow d_i = -1$$

are LINEARLY SEPARABLE in $\mathbb{R}^p$?



NOT LINEARLY SEPARABLE

LINEARLY SEPARABLE

usually $\boxed{p \gg n}$

(Or $p = \infty$ is also possible)

- Usually possible (if $p \gg n$)
- How to choose $\phi(\cdot)$?
- Use typical ~~basis~~ functions (polynomials, gaussian terms, wavelets, etc.)

<u>EXAMPLE:</u>

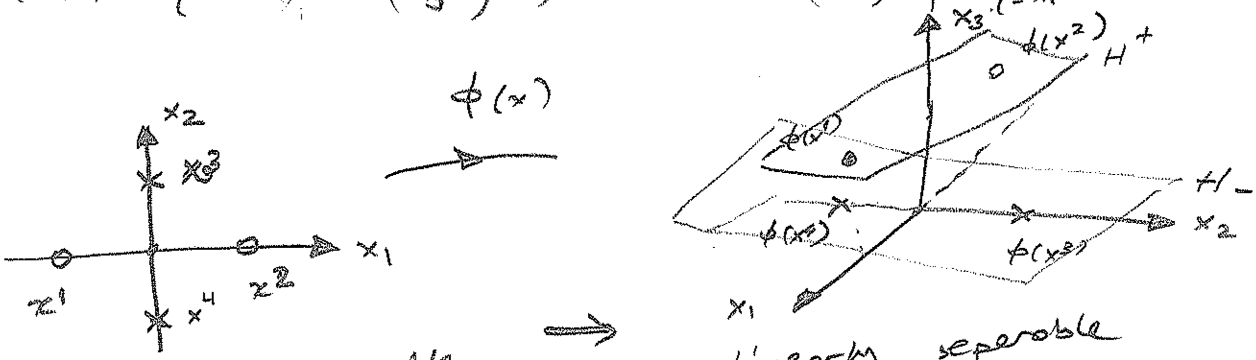
$$C_1 = \left\{ x^1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \; x^2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} \quad d_i = 1$$

$$C_2 = \left\{ x^3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \; x^4 = \begin{pmatrix} 0 \\ -1 \end{pmatrix} \right\} \quad d_i = -1$$

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2, \text{ use } \quad \phi(x) = \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 \end{pmatrix} \in \mathbb{R}^3$$

$$\phi(C_1) = \left\{ \phi(x^1) = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \; \phi(x^2) = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \right\} \quad d_i = 1$$

$$\phi(C_2) = \left\{ \phi(x^3) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \; \phi(x^4) = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix} \right\} \quad d_i = -1$$



Not linearly separable $\Longrightarrow$ Linearly separable

$H^+ : \; x_3 - 1 = 0$

$H^- : \; x_3 = 0$

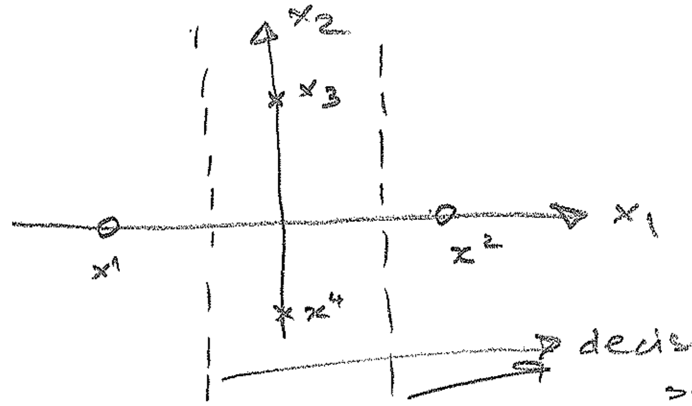
$x_1^2 - 1/2 = 0 \quad \Longleftarrow \quad H : \; x_3 - \frac{1}{2} = 0$

Discriminant function: $\quad g(x) = x_1^2 - 1/2$

$$\operatorname{sgn}[g(x)] \begin{cases} \nearrow +1 \iff x \in C_1 \\ \searrow -1 \iff x \in C_2 \end{cases}$$

- $\phi(x)$ space is called feature space
- Optimal plane in feature space
- May not be optimal in original space.

$$g(x) = x_1^2 - 1/2 \Rightarrow g(x) = 0 \Rightarrow x_1 = \mp 1/\sqrt{2}$$



decision bdry in original space

$x_1 = -1/\sqrt{2}$ $\quad$ $x_1 = 1/\sqrt{2}$ $\quad$ Not in the middle!

- How to formulate?

Remember dual problem:

$$\boxed{L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j x_i^T x_j}$$

$$\boxed{\begin{array}{lll} DP & \max_{\alpha} & L(\alpha) \\ & \text{s.t.} & \alpha_i \geq 0 \\ & & \sum_{i=1}^{N} \alpha_i d_i = 0 \end{array}}$$

$$\boxed{g(x) = \sum_{i=1}^{N} \alpha_i d_i x_i - \theta}$$

$$\boxed{\theta_{ave} = \frac{1}{N_s} \sum_{i=1}^{N_s} (w^T x_i - d_i)}$$

N_s: support vectors.

— Generalization to Nonlinear case:

replace $x_i \Rightarrow \phi(x_i)$

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \phi(x_i)^T \phi(x_j)$$

DP. max $L(\alpha)$

s.t. $\alpha_i \geq 0$

$\sum_{i=1}^{N} \alpha_i d_i = 0$

$$\boxed{g(x) = w^T \phi(x) - \theta}$$

$$\boxed{w = \sum_{i=1}^{N} \alpha_i d_i \phi(x_i)}$$

$$\boxed{\theta = \frac{1}{N_s} \sum_{i=1}^{N_s} \left(w^T \phi(x_i) - d_i \right)}$$

N_s: support vectors

KERNEL TRICK

$$\boxed{K(x_i, x_j) = \phi(x_i)^T \phi(x_j)}$$

— Above formulation works with $K(x_i, x_j)$

— $K(x_i, x_j)$ is called a KERNEL

$\rightarrow$

$$g(x) = \sum_{i=1}^{N} \alpha_i d_i \phi(x_i)^T \phi(x) - \theta$$

$$\boxed{g(x) = \sum_{i=1}^{N} \alpha_i d_i K(x_i, x) - \theta}$$

Possible kernels.

— Linear : $K(x_i, x_j) = x_i^T x_j \Rightarrow$ Linear SVM

— Polynomial : $K(x_i, x_j) = (1 + x_i^T x_j)^d$ (d: degree, an integer.

— Gaussian : $K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{\sigma^2}}$

— General : $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$

Remember: $g(x) = \sum_{i=1}^{N} \alpha_i d_i \phi(x_i)^T \phi(x) - \theta$

— Only support vectors have $\alpha_i > 0$.