IIC 2433 Minería de Datos

https://github.com/marcelomendoza/IIC2433

- SUPPORT VECTOR MACHINES -

SVM en datos no separables (linealmente)

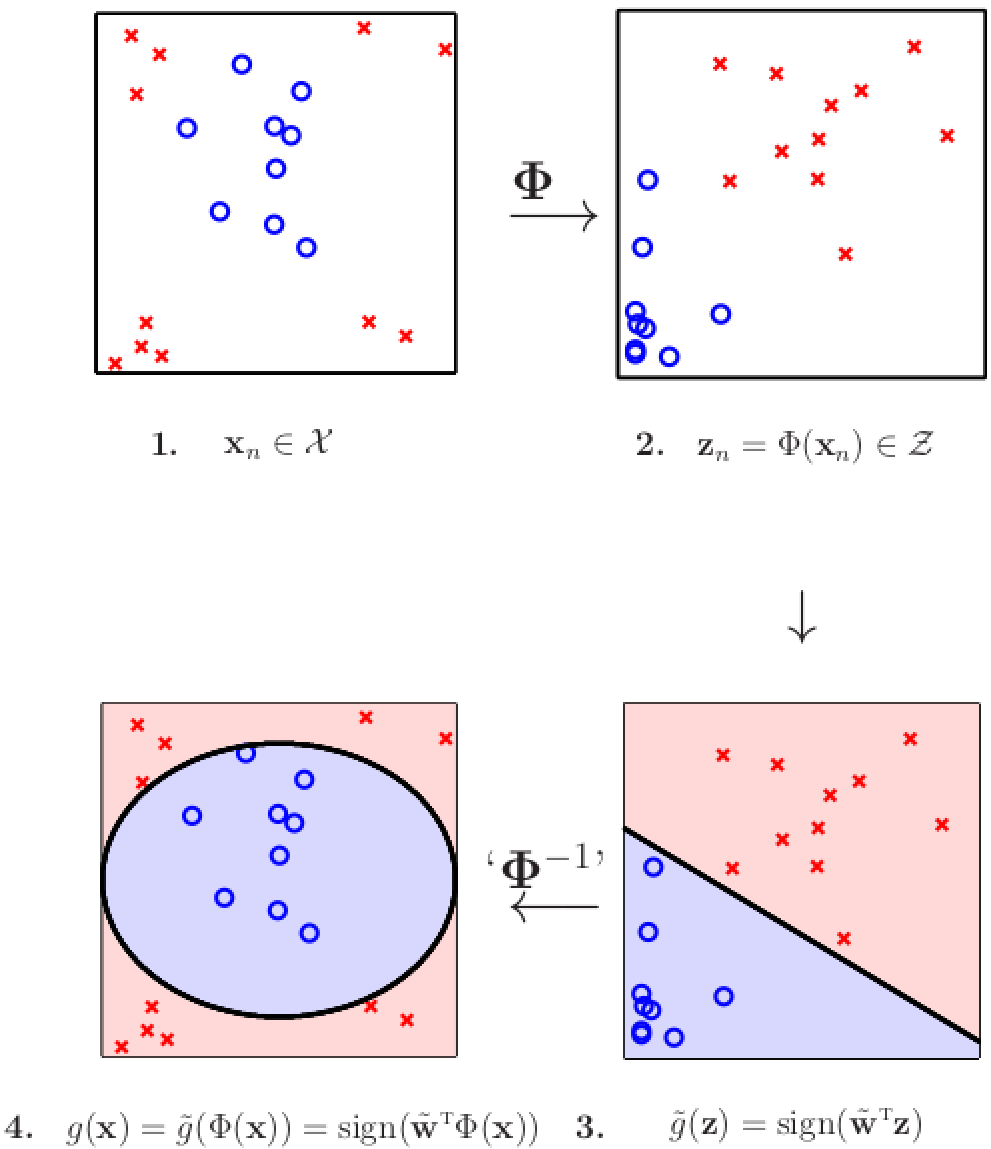
Consideremos una transformación $\Phi : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ tal que $\mathbf{z}_n = \mathbf{\Phi}(\mathbf{x}_n)$.
Después de transformar los datos, el problema SVM (hard margin) es:

$$\begin{array}{lll} \underset{\tilde{b}, \tilde{\mathbf{w}}}{\text{minimize:}} & \frac{1}{2}\tilde{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{w}} & \\ \text{subject to:} & y_n\left(\tilde{\mathbf{w}}^{\mathrm{T}}\mathbf{z}_n + \tilde{b}\right) \geq 1 & (n = 1, \cdots, N), \end{array}$$

donde $\tilde{\mathbf{w}}$ reside en el espacio de representación *Z*. Notemos que la dimensionalidad de Z puede ser distinta a la de la entrada.

Si la transformación es no lineal, puede ayudar a la SVM a separar datos no separables en el espacio original.

SVM en datos no separables (linealmente)



$$\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix} \qquad \mathbf{z} = \mathbf{\Phi}(\mathbf{x}) = \begin{bmatrix} 1 \\ \Phi_1(\mathbf{x}) \\ \vdots \\ \Phi_{\tilde{d}}(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} 1 \\ z_1 \\ \vdots \\ z_{\tilde{d}} \end{bmatrix}$$

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ $\qquad$ $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N$

$y_1, y_2, \ldots, y_N$ $\qquad$ $y_1, y_2, \ldots, y_N$

$$g(\mathbf{x}) = \text{sign}(\tilde{\mathbf{w}}^{\mathrm{T}}\mathbf{\Phi}(\mathbf{x})) \qquad \tilde{\mathbf{w}} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_{\tilde{d}} \end{bmatrix}$$

Este problema es difícil de resolver cuando $\tilde{d}$ es muy grande.

SVM en datos no separables (linealmente)

En la práctica, en lugar de trabajar sobre el problema original en el espacio *Z*, se aborda la formulación dual ya que es más fácil desde el punto de vista de optimización.

Teorema (KKT). El problema QP convexo en su forma primal:

$$\begin{array}{ll} \underset{\mathbf{u}\in\mathbb{R}^L}{\text{minimize:}} & \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathrm{Q}\mathbf{u} + \mathbf{p}^{\mathrm{T}}\mathbf{u} \\ \text{subject to:} & \mathbf{a}_m^{\mathrm{T}}\mathbf{u} \geq c_m \end{array}$$

define la función dual (Lagrange):

$$\mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathrm{Q}\mathbf{u} + \mathbf{p}^{\mathrm{T}}\mathbf{u} + \sum_{m=1}^{M} \alpha_m \left(c_m - \mathbf{a}_m^{\mathrm{T}}\mathbf{u}\right).$$

La solución del primal es óptima ssi es óptima en el dual. Luego:

$$\max_{\boldsymbol{\alpha}\geq\mathbf{0}} \min_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}).$$

$$\max_{\boldsymbol{\alpha} \geq \mathbf{0}} \min_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}).$$

$$\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \in \mathbb{R}^{d+1}$$

SVM en datos no separables (linealmente)

Apliquemos la formulación dual a la SVM:

$$\begin{aligned} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) &= \tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \sum_{n=1}^{N} \alpha_n \left(1 - y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b)\right) \\ &= \tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{w}^{\mathrm{T}}\mathbf{x}_n - b\sum_{n=1}^{N} \alpha_n y_n + \sum_{n=1}^{N} \alpha_n \end{aligned}$$

$$\max_{\boldsymbol{\alpha} \geq \mathbf{0}} \min_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}).$$

$$\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \in \mathbb{R}^{d+1}$$

SVM en datos no separables (linealmente)

Apliquemos la formulación dual a la SVM:

$$\begin{aligned} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) &= \tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \sum_{n=1}^{N} \alpha_n \left(1 - y_n(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n + b)\right) \\ &= \tfrac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{w}^{\mathrm{T}}\mathbf{x}_n - b\sum_{n=1}^{N} \alpha_n y_n + \sum_{n=1}^{N} \alpha_n \end{aligned}$$

Minimizamos w.r.t. $(b, \mathbf{w})$

$\frac{\partial \mathcal{L}}{\partial b} = 0$:

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{n=1}^{N} \alpha_n y_n \quad \Longrightarrow \quad \sum_{n=1}^{N} \alpha_n y_n = 0$$

$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0$:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n \quad \Longrightarrow \quad \mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$$

SVM en datos no separables (linealmente)

Usaremos $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$ en $\frac{1}{2}\mathbf{w}^\mathrm{T}\mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{w}^\mathrm{T}\mathbf{x}_n$

$$\begin{aligned}
\mathcal{L} &= \frac{1}{2}\mathbf{w}^\mathrm{T}\mathbf{w} - \mathbf{w}^\mathrm{T}\sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n - b\underbrace{\sum_{n=1}^{N} \alpha_n y_n}_{0} + \sum_{n=1}^{N} \alpha_n \\
&= -\frac{1}{2}\mathbf{w}^\mathrm{T}\mathbf{w} + \sum_{n=1}^{N} \alpha_n \\
&= -\frac{1}{2}\sum_{m,n=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{x}_n^\mathrm{T}\mathbf{x}_m + \sum_{n=1}^{N} \alpha_n
\end{aligned}$$

SVM en datos no separables (linealmente)

Usaremos $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$ en $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{w}^{\mathrm{T}}\mathbf{x}_n + \sum_{n=1}^{N} \alpha_n.$

0

$$\begin{aligned}
\mathcal{L} &= \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n - b\cancel{\sum_{n=1}^{N} \alpha_n y_n} + \sum_{n=1}^{N} \alpha_n \\
&= -\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \sum_{n=1}^{N} \alpha_n \\
&= -\frac{1}{2}\sum_{m,n=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_m + \sum_{n=1}^{N} \alpha_n
\end{aligned}$$

Ojo que:

$$\max_{\boldsymbol{\alpha} \geq \mathbf{0}} \min_{\mathbf{u}} \mathcal{L}(\mathbf{u}, \boldsymbol{\alpha}).$$

es lo mismo que

Es decir, debemos hacer lo siguiente:

$$\underset{\boldsymbol{\alpha} \in \mathbb{R}^N}{\text{minimize}}: \quad \frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N} y_n y_m \alpha_n \alpha_m \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_m - \sum_{n=1}^{N} \alpha_n$$

$$\text{subject to:} \quad \sum_{n=1}^{N} y_n \alpha_n = 0$$

$$\alpha_n \geq 0 \qquad (n = 1, \cdots, N).$$

SVM en datos no separables (linealmente)

El dual se puede reescribir en una forma más amable:

$$\begin{aligned}
\underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\text{minimize}}: \quad & \frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N} y_n y_m \alpha_n \alpha_m \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_m - \sum_{n=1}^{N}\alpha_n \\
\text{subject to:} \quad & \sum_{n=1}^{N} y_n \alpha_n = 0 \\
& \alpha_n \geq 0 \qquad\qquad (n = 1, \cdots, N).
\end{aligned}$$

$$\begin{aligned}
\underset{\boldsymbol{\alpha}}{\text{minimize}} \quad & \tfrac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{G}\boldsymbol{\alpha} - \mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha} \\
\text{subject to:} \quad & \mathbf{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0 \\
& \boldsymbol{\alpha} \geq \mathbf{0}
\end{aligned}$$

donde $(\mathrm{G}_{nm} = y_n y_m \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_m)$

SVM en datos no separables (linealmente)

Podemos recuperar la SVM del primal desde el dual:

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n^* y_n \mathbf{x}_n$$

$$\begin{aligned} \alpha_s > 0 \quad & \Longrightarrow \quad y_s(\mathbf{w}^{\mathrm{T}}\mathbf{x}_s + b) - 1 = 0 \\ & \Longrightarrow \quad b = y_s - \mathbf{w}^{\mathrm{T}}\mathbf{x}_s \end{aligned}$$

SVM en datos no separables (linealmente)

Podemos recuperar la SVM del primal desde el dual:

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n^* y_n \mathbf{x}_n$$

$$\alpha_s > 0 \quad \Longrightarrow \quad y_s(\mathbf{w}^{\mathrm{T}}\mathbf{x}_s + b) - 1 = 0$$

$$\Longrightarrow \quad b = y_s - \mathbf{w}^{\mathrm{T}}\mathbf{x}_s$$

$(\mathrm{G}_{nm} = y_n y_m \mathbf{x}_n^{\mathrm{T}} \mathbf{x}_m)$

Una forma simple de calcular G:

signed data matrix ↓

$$\mathrm{X} = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix} \quad \longrightarrow \quad \mathrm{X}_s = \begin{bmatrix} 0 & 0 \\ -2 & -2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \longrightarrow \quad \mathrm{G} = \mathrm{X}_s \mathrm{X}_s^{\mathrm{T}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 8 & -4 & -6 \\ 0 & -4 & 4 & 6 \\ 0 & -6 & 6 & 9 \end{bmatrix}$$

SVM en datos no separables (linealmente)

$$\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \in \mathbb{R}^{d+1}$$

<u>QP problem</u>

$$\begin{array}{ll} \underset{\mathbf{u}}{\text{minimize}} & \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathrm{Q}\mathbf{u} + \mathbf{p}^{\mathrm{T}}\mathbf{z} \\ \text{subject to:} & \mathrm{A}\mathbf{u} \geq \mathbf{c} \end{array}$$

<u>Dual SVM</u>

$$\begin{array}{ll} \underset{\boldsymbol{\alpha}}{\text{minimize}} & \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{G}\boldsymbol{\alpha} - \mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha} \\ \text{subject to:} & \mathbf{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0 \\ & \boldsymbol{\alpha} \geq \mathbf{0} \end{array}$$

$$\left.\begin{array}{r} \mathbf{u} = \boldsymbol{\alpha} \\ \mathrm{Q} = \mathrm{G} \\ \mathbf{p} = -\mathbf{1}_N \\ \mathrm{A} = \begin{bmatrix} \mathbf{y}^{\mathrm{T}} \\ -\mathbf{y}^{\mathrm{T}} \\ \mathrm{I}_N \end{bmatrix} \\ \mathbf{c} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{0}_N \end{bmatrix} \end{array}\right\} \xrightarrow{\mathrm{QP}(\mathrm{Q},\mathbf{p},\mathrm{A},\mathbf{c})}$$

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 1 \\ 0 \end{bmatrix}$$

$$\mathbf{w} = \sum_{n=1}^{4} \alpha_n^* y_n \mathbf{x}_n = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$b = y_1 - \mathbf{w}^{\mathrm{T}}\mathbf{x}_1 = -1$$

$$\gamma = \frac{1}{\|\mathbf{w}\|} = \frac{1}{\sqrt{2}}$$

SVM en datos no separables (linealmente)

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n^* y_n \mathbf{x}_n$$

$$\begin{aligned} \alpha_s > 0 \quad &\Longrightarrow \quad y_s(\mathbf{w}^{\mathrm{T}}\mathbf{x}_s + b) - 1 = 0 \\ &\Longrightarrow \quad b = y_s - \mathbf{w}^{\mathrm{T}}\mathbf{x}_s \end{aligned}$$

QP

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 1 \\ 0 \end{bmatrix}$$

$$\mathbf{w} = \sum_{n=1}^{4} \alpha_n^* y_n \mathbf{x}_n = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$b = y_1 - \mathbf{w}^{\mathrm{T}}\mathbf{x}_1 = -1$$

$$\gamma = \frac{1}{\|\mathbf{w}\|} = \frac{1}{\sqrt{2}}$$

SVM en datos no separables (linealmente)

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n^* y_n \mathbf{x}_n$$

$$\alpha_s > 0 \implies y_s(\mathbf{w}^{\mathrm{T}}\mathbf{x}_s + b) - 1 = 0$$
$$\implies b = y_s - \mathbf{w}^{\mathrm{T}}\mathbf{x}_s$$

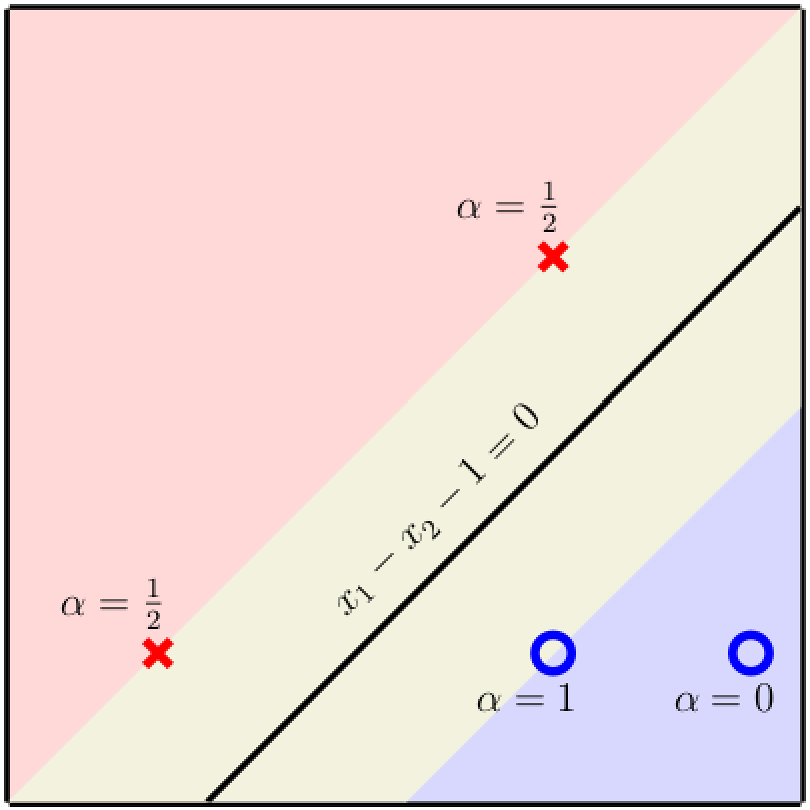
QP

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 1 \\ 0 \end{bmatrix}$$

$$\mathbf{w} = \sum_{n=1}^{4} \alpha_n^* y_n \mathbf{x}_n = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$b = y_1 - \mathbf{w}^{\mathrm{T}}\mathbf{x}_1 = -1$$

$$\gamma = \frac{1}{\|\mathbf{w}\|} = \frac{1}{\sqrt{2}}$$



Los que no son vectores de soporte tienen: $\alpha_n = 0$

SVM en datos no separables (linealmente)

Resolver el problema en el dual nos permite trabajar en el espacio *Z.*

$$\underset{\boldsymbol{\alpha}}{\text{minimize}} \quad \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{G}\boldsymbol{\alpha} - \mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha}$$

$$\text{subject to: } \mathbf{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0$$

$$\mathbf{C} \geq \boldsymbol{\alpha} \geq \mathbf{0}$$

$$G_{nm} = y_n y_m (\mathbf{z}_n^{\mathrm{T}} \mathbf{z}_m)$$

$$g(\mathbf{x}) = \operatorname{sign}\left(\sum_{\alpha_n^* > 0} \alpha_n^* y_n (\mathbf{z}_n^{\mathrm{T}} \mathbf{z}) + b^* \right)$$

$$C > \alpha_s^* > 0$$

$$b^* = y_s - \sum_{\alpha_n^* > 0} \alpha_n^* y_n (\mathbf{z}_n^{\mathrm{T}} \mathbf{z}_s)$$

producto interno

SVM en datos no separables (linealmente)

Un kernel es una función que combina tanto la transformación como el producto interno:

$$K_\Phi(\mathbf{x}, \mathbf{x}') \equiv \Phi(\mathbf{x})^{\mathrm{T}}\Phi(\mathbf{x}').$$

El kernel toma dos vectores y entrega el producto interno en *Z*.

SVM en datos no separables (linealmente)

Un kernel es una función que combina tanto la transformación como el producto interno:

$$K_{\Phi}(\mathbf{x}, \mathbf{x}') \equiv \Phi(\mathbf{x})^{\mathrm{T}} \Phi(\mathbf{x}').$$

El kernel toma dos vectores y entrega el producto interno en *Z*.

Ejemplo: Kernel polinomial de segundo orden.

$$\Phi_2(\mathbf{x})^{\mathrm{T}} \Phi_2(\mathbf{x}') \quad = \quad 1 + (\mathbf{x}^{\mathrm{T}} \mathbf{x}') + (\mathbf{x}^{\mathrm{T}} \mathbf{x}')^2.$$

SVM en datos no separables (linealmente)

1: **Input:** $\mathrm{X}, \mathbf{y}$

2: Compute G: $\mathrm{G}_{nm} = y_n y_m K(\mathbf{x}_n, \mathbf{x}_m)$.

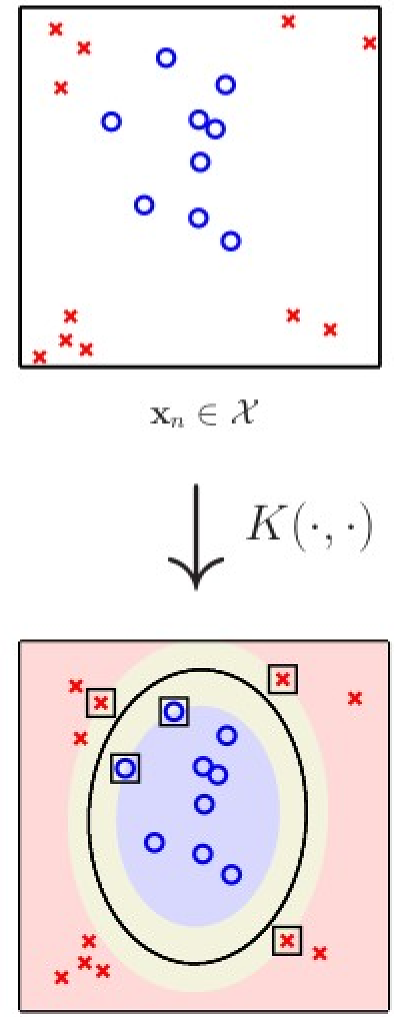
3: Solve (QP):

$$\left.\begin{array}{ll} \underset{\boldsymbol{\alpha}}{\text{minimize:}} & \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathrm{G}\boldsymbol{\alpha} - \mathbf{1}^{\mathrm{T}}\boldsymbol{\alpha} \\ \text{subject to:} & \mathbf{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0 \\ & \boldsymbol{\alpha} \geq \mathbf{0} \end{array}\right\} \longrightarrow \begin{array}{l} \boldsymbol{\alpha}^* \\ \text{index } s : \quad \alpha_s^* > 0 \end{array}$$

4: $$b^* = y_s - \sum_{\alpha_n^* > 0} \alpha_n^* y_n K(\mathbf{x}_n, \mathbf{x}_s)$$

5: The final hypothesis is

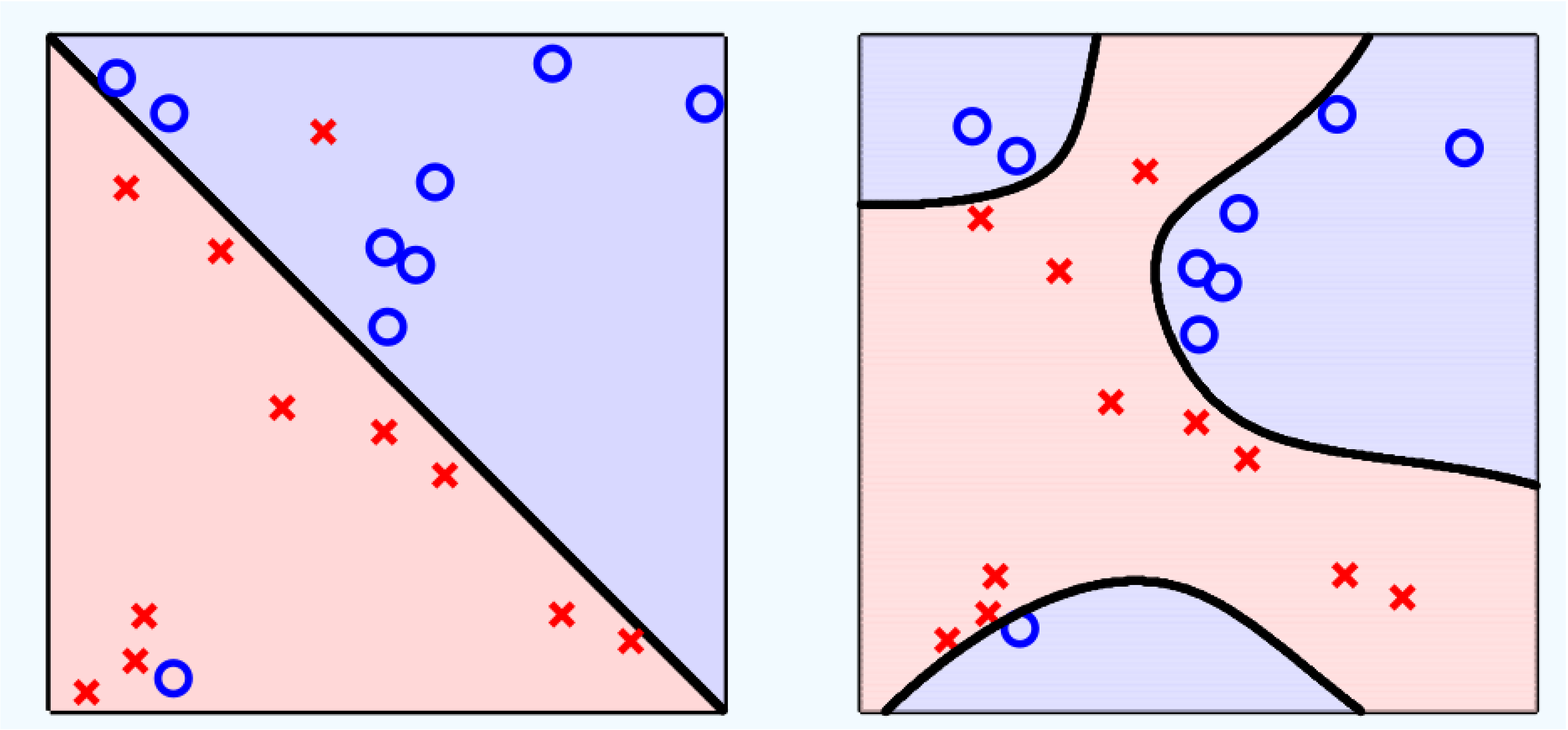
$$g(\mathbf{x}) = \mathrm{sign}\left(\sum_{\alpha_n^* > 0} \alpha_n^* y_n K(\mathbf{x}_n, \mathbf{x}) + b^*\right)$$



$$g(\mathbf{x}) = \mathrm{sign}\left(\sum_{\alpha_n^* > 0} \alpha_n^* y_n K(\mathbf{x}_n, \mathbf{x}) + b^*\right)$$

$$b^* = y_s - \sum_{\alpha_n^* > 0} \alpha_n^* y_n K(\mathbf{x}_n, \mathbf{x}_s)$$

SVM con kernel Gaussiano



a) SVM lineal   b) SVM con kernel Gaussiano