Aprendizaje de máquinas
Support vector machines (parte 2)

Felipe Tobar

Facultad de Ciencias Físicas y Matemáticas
Universidad de Chile

Otoño, 2021.

Método del kernel: motivación

Si bien SVM tiene buenas propiedades de generalización, el método se cae cuando los datos no son linealmente separables.

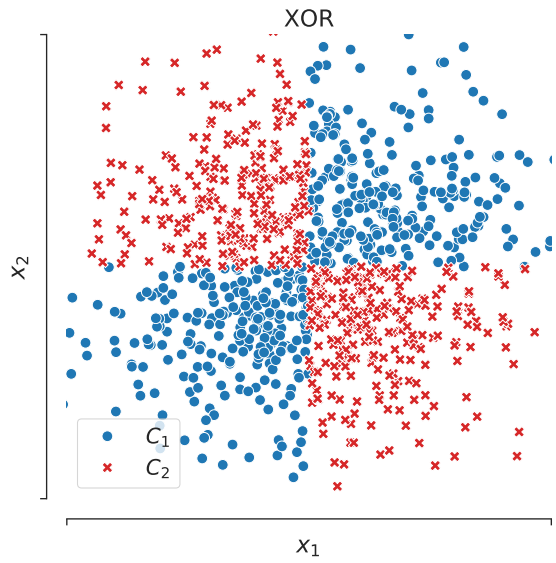
Método del kernel: motivación

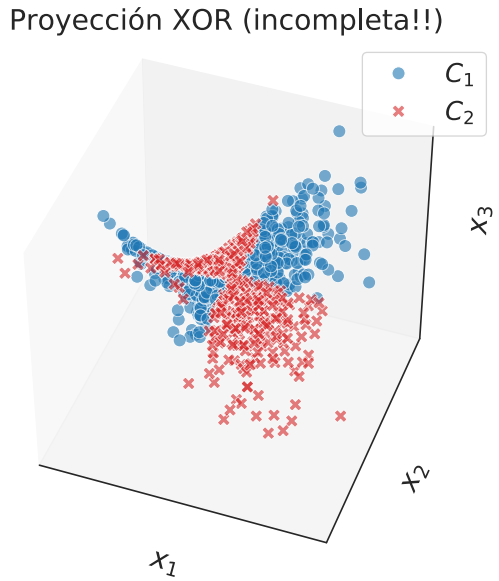
Si bien SVM tiene buenas propiedades de generalización, el método se cae cuando los datos no son linealmente separables.

Sin embargo, para la figura de la izquierda, es posible diseñar una característica particular donde los datos sí son linealmente separables. En efecto, consideremos el mapa desde $\mathbb{R}^2$ a $\mathbb{R}^3$ definido mediante

$$\phi : [x_1, x_2]^\top \mapsto [x_1, x_2, x_1 x_2]^\top$$

Esta nueva representación permite separar de forma lineal las clases mediante el plano $z = 0$ en $\mathbb{R}^3$ tal como se observa en la figura derecha.





Método del kernel: idea general

En el caso general no es claro cuál debe ser "el buen ϕ". A pesar de esto, notemos que en la formulación de SVM solo se requiere poder calcular los productos internos entre las características de cada entrada, es decir, si se considera un ϕ arbitrario para el problema de clasificación, solo se necesitaría calcular los productos internos de la forma

$$\langle \phi(x_i), \phi(x_j) \rangle$$

Método del kernel: idea general

En el caso general no es claro cuál debe ser "el buen ϕ". A pesar de esto, notemos que en la formulación de SVM solo se requiere poder calcular los productos internos entre las características de cada entrada, es decir, si se considera un ϕ arbitrario para el problema de clasificación, solo se necesitaría calcular los productos internos de la forma

$$\langle \phi(x_i), \phi(x_j) \rangle$$

Por lo tanto, si bien no siempre se puede encontrar el ϕ exclusivo de cada problema, se puede utilizar uno que sea muy general, esperando que alguno de los términos agregados sea el que efectivamente separa las clases.

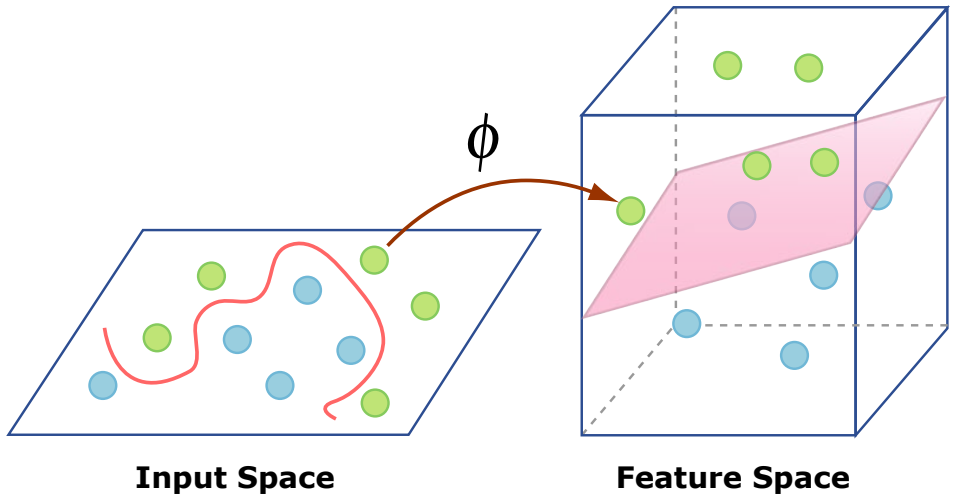
Método del kernel: idea general

En el caso general no es claro cuál debe ser "el buen ϕ". A pesar de esto, notemos que en la formulación de SVM solo se requiere poder calcular los productos internos entre las características de cada entrada, es decir, si se considera un ϕ arbitrario para el problema de clasificación, solo se necesitaría calcular los productos internos de la forma

$$\langle \phi(x_i), \phi(x_j) \rangle$$

Por lo tanto, si bien no siempre se puede encontrar el ϕ exclusivo de cada problema, se puede utilizar uno que sea muy general, esperando que alguno de los términos agregados sea el que efectivamente separa las clases.

De esta forma, se puede considerar un mapeo de alta dimensión que incorpore varias combinaciones entre las componentes.



Método del kernel: Mercer kernel

Para encontrar estos ϕ generales, veamos la siguiente definición:

Definition (Mercer kernel)

Un Mercer kernel es una función continua $K : X \times X \to \mathbb{R}$ tal que

- Es simétrica $K(x_1, x_2) = K(x_2, x_1)$
- Es definida positiva, es decir

$$\int_{X^2} K(x_1, x_2) g(x_1) g(x_2) dx_1 dx_2 \geq 0,$$

para toda función $g : X \to \mathbb{R}$ continua.

Método del kernel: Mercer kernel

Para encontrar estos ϕ generales, veamos la siguiente definición:

Definition (Mercer kernel)

Un Mercer kernel es una función continua $K : X \times X \to \mathbb{R}$ tal que

- Es simétrica $K(x_1, x_2) = K(x_2, x_1)$
- Es definida positiva, es decir

$$\int_{X^2} K(x_1, x_2) g(x_1) g(x_2) dx_1 dx_2 \geq 0,$$

para toda función $g : X \to \mathbb{R}$ continua.

El nombre de la segunda propiedad derivada de su similitud con las matrices definidas positivas: si g se mira como vector de $\mathbb{R}^X$, entonces la expresión anterior representa a $g^\top K g \leq 0$.

Método del kernel: teorema de Mercer

De esta forma, se tiene el siguiente teorema de analisis funcional que sustenta la utilización de kernels en los algoritmos de aprendizaje automático:

Theorem (teorema de Mercer (simplificado))

Sea $K : X \times X \to \mathbb{R}$ un Mercer kernel, entonces existe un espacio de Hilbert $(\mathcal{H}, \langle , \rangle)$ y una función $\phi : X \to \mathcal{H}$ tal que:

$$K(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$$

Método del kernel: teorema de Mercer

De esta forma, se tiene el siguiente teorema de analisis funcional que sustenta la utilización de kernels en los algoritmos de aprendizaje automático:

Theorem (teorema de Mercer (simplificado))

Sea $K : X \times X \to \mathbb{R}$ un Mercer kernel, entonces existe un espacio de Hilbert $(\mathcal{H}, \langle , \rangle)$ y una función $\phi : X \to \mathcal{H}$ tal que:

$$K(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$$

Es decir, existe un mapa de características ϕ tal que $K(x_1, x_2)$ representa el producto interno (en algún espacio) de las características de x_1 y x_2. Además, dicho espacio no es necesariamente de dimensión finita.

Ejemplos de kernels

Kernel polinomial: este kernel está definido por

$$K_{pol}(x, y) = (c + x^\top y)^d$$

donde $c \geq 0$ es un parámetro libre y $d \in \mathbb{N}$ es el orden del polinomio.

Ejemplos de kernels

Kernel polinomial: este kernel está definido por

$$K_{pol}(x, y) = (c + x^\top y)^d$$

donde $c \geq 0$ es un parámetro libre y $d \in \mathbb{N}$ es el orden del polinomio. Para el caso $d = 2$ se pueden reagrupar los términos para ver que el mapa de características que induce este kernel es

$$\phi_{pol}(x) = [x_1^2, \ldots, x_m^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_mx_{m-1}, \sqrt{2c}x_1, \ldots, \sqrt{2c}x_m, c].$$

Ejemplos de kernels

Kernel polinomial: este kernel está definido por

$$K_{pol}(x, y) = (c + x^\top y)^d$$

donde $c \geq 0$ es un parámetro libre y $d \in \mathbb{N}$ es el orden del polinomio. Para el caso $d = 2$ se pueden reagrupar los términos para ver que el mapa de características que induce este kernel es

$$\phi_{pol}(x) = [x_1^2, \ldots, x_m^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_mx_{m-1}, \sqrt{2c}x_1, \ldots, \sqrt{2c}x_m, c].$$

Es decir, al usar el kernel poliomial se está usando un mapa de características que contiene todos los monomios de grado hasta $d = 2$. Esta propiedad se cumple en general para cualquier $d \in \mathbb{N}$.

Ejemplos de kernels

Función de base radial (RBF kernel): este kernel también se denomina kernel exponencial cuadrático o gaussiano, y está definido por

$$K_{RBF}(x, y) = \sigma^2 \exp\left(-\frac{\|x - y\|^2}{2l^2}\right)$$

El mapa de características que induce es de dimensión infinita y las fronteras que entrega son suaves (infinitamente diferenciables).

Ejemplos de kernels

Función de base radial (RBF kernel): este kernel también se denomina kernel exponencial cuadrático o gaussiano, y está definido por

$$K_{RBF}(x, y) = \sigma^2 \exp\left(-\frac{\|x - y\|^2}{2l^2}\right)$$

El mapa de características que induce es de dimensión infinita y las fronteras que entrega son suaves (infinitamente diferenciables).

Kernel periódico: está definido como

$$K_{per}(x, y) = \sigma^2 \exp\left(-\frac{2\,\mathrm{sen}^2\left(\frac{\pi|x-y|}{p}\right)}{l^2}\right).$$

Este kernel es capaz de rescatar características periódicas en los datos (controlados por el parámetro p).

Truco del kernel

La introducción del concepto de kernel es fundamental en SVM ya que:

- El teorema de Mercer afirma que para cualquier función K simétrica y definida positiva, $K(x_1, x_2)$ representa un producto interno en algún espacio de características.

Truco del kernel

La introducción del concepto de kernel es fundamental en SVM ya que:

- El teorema de Mercer afirma que para cualquier función K simétrica y definida positiva, $K(x_1, x_2)$ representa un producto interno en algún espacio de características.
- Dado que SVM solo requiere del cálculo de productos internos, lo anterior permite construir *kernel SVMs*, donde se parametriza directamente el producto interno en el problema de optimización mediante el kernel, pues el teorema de Mercer da la garantía que el mapa de características ϕ existe.

Truco del kernel

La introducción del concepto de kernel es fundamental en SVM ya que:

- El teorema de Mercer afirma que para cualquier función K simétrica y definida positiva, $K(x_1, x_2)$ representa un producto interno en algún espacio de características.
- Dado que SVM solo requiere del cálculo de productos internos, lo anterior permite construir *kernel SVMs*, donde se parametriza directamente el producto interno en el problema de optimización mediante el kernel, pues el teorema de Mercer da la garantía que el mapa de características ϕ existe.
- Este truco puede aplicarse a cualquier algoritmo en donde las entradas solo aparezcan en la forma de productos punto, proceso que recibe el nombre de *kernelización*.

Kernel SVM

Dado un mapa de características ϕ, en la formulación de SVM se pueden reemplazar las entadas por las características inducidas por ϕ. De este modo, el problema primal es:

$$
\begin{aligned}
(P) \quad & \min_{w,b} \quad \frac{1}{2}||w||^2 + c\sum_{i=1}^{N}\xi_i \\
& \text{s.a} \quad y_i(w^\top \phi(x_i) + b) \geq 1 - \xi_i,\ i = 1, \ldots, N.
\end{aligned}
$$

Kernel SVM

Dado un mapa de características ϕ, en la formulación de SVM se pueden reemplazar las entadas por las características inducidas por ϕ. De este modo, el problema primal es:

$$
\begin{aligned}
(P) \quad \min_{w,b} \quad & \frac{1}{2}||w||^2 + c\sum_{i=1}^{N}\xi_i \\
\text{s.a} \quad & y_i(w^\top \phi(x_i) + b) \geq 1 - \xi_i,\ i = 1, \ldots, N.
\end{aligned}
$$

Mientras que su formulación dual tiene la forma

$$
\begin{aligned}
(D) \quad \max_{\alpha} \quad & \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\alpha_i\alpha_j y_i y_j \langle \phi(x_i), \phi(x_j)\rangle \\
\text{s.a} \quad & \sum_{i=1}^{N}\alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq c
\end{aligned}
$$

Kernel SVM

Dado que el problema de optimización solo utiliza los datos mediante productos internos, se puede ocupar el truco del kernel para parametrizar directamente el producto interno $\langle \phi(x_i), \phi(x_j) \rangle$ mediante $K(x_i, x_j)$.

Kernel SVM

Dado que el problema de optimización solo utiliza los datos mediante productos internos, se puede ocupar el truco del kernel para parametrizar directamente el producto interno $\langle \phi(x_i), \phi(x_j) \rangle$ mediante $K(x_i, x_j)$.

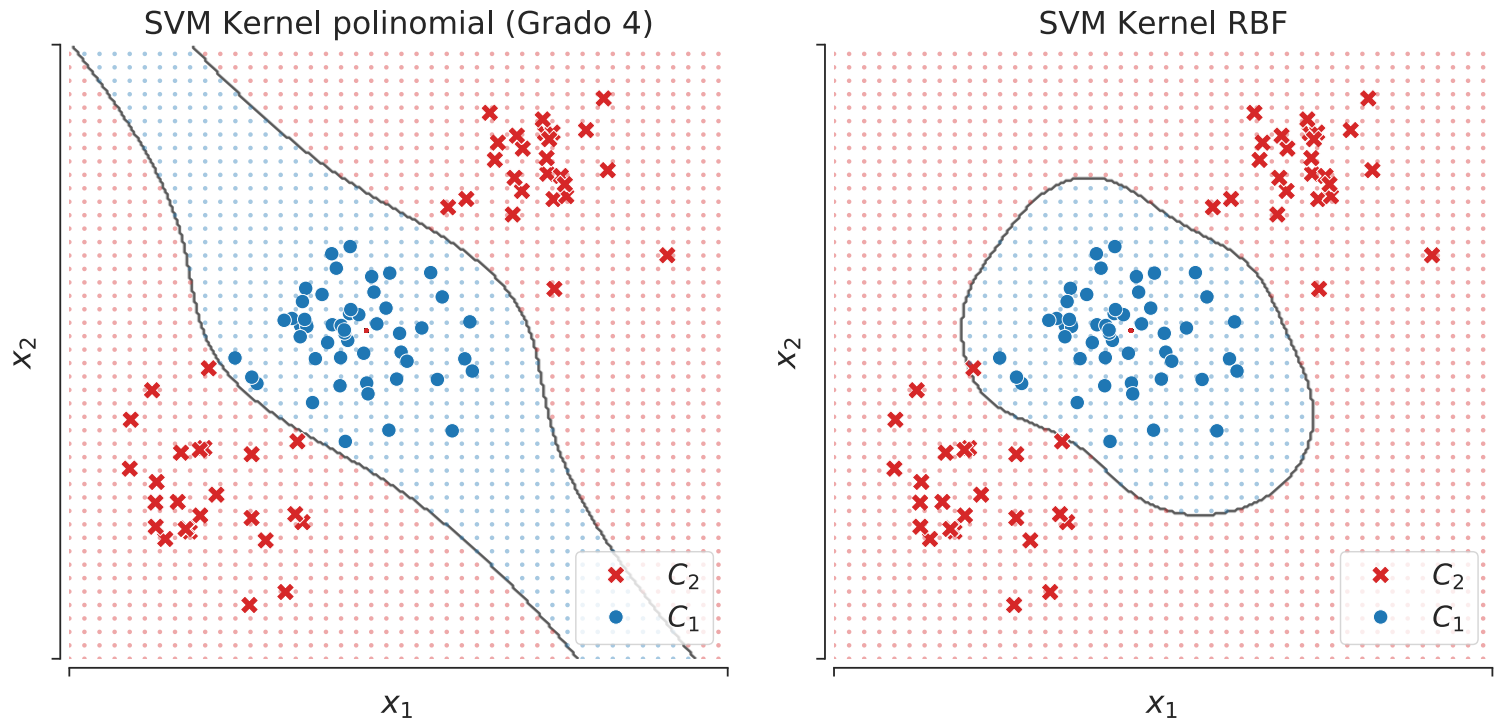
Con esto, el problema de optimización en el dual se convierte en

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{s.a} \quad & \sum_{i=1}^{N} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq c \end{aligned}$$

Kernel SVM: ejemplo

La siguiente figura muestra la implementación de kernel SVM para dos kernels:

- A la izquierda se utilizó un kernel polinomial de grado 3
- A la derecha se utilizó un kernel un kernel RBF.



Kernel ridge regression

Otro método en el que se puede utilizar el truco del kernel es el método de regulación cuadrática usada en regresión lineal, el cual tiene solución

$$\theta_{MCR} = \left(\tilde{X}^\top \tilde{X} + \rho \mathbb{I}\right)^{-1} \tilde{X}^\top Y$$

Kernel ridge regression

Otro método en el que se puede utilizar el truco del kernel es el método de regulación cuadrática usada en regresión lineal, el cual tiene solución

$$\theta_{MCR} = \left(\tilde{X}^\top \tilde{X} + \rho \mathbb{I}\right)^{-1} \tilde{X}^\top Y$$

Si bien las entradas $\tilde{x}_i$ no aparecen en forma de producto interno, es posible reescribir la solución de otra forma mediante la fórmula de Woodburry. De este modo se obtiene que

$$\theta_{MCR} = \tilde{X}^\top \left(\tilde{X}\tilde{X}^\top + \rho \mathbb{I}\right)^{-1} Y$$

Kernel ridge regression

Otro método en el que se puede utilizar el truco del kernel es el método de regulación cuadrática usada en regresión lineal, el cual tiene solución

$$\theta_{MCR} = \left(\tilde{X}^\top \tilde{X} + \rho \mathbb{I}\right)^{-1} \tilde{X}^\top Y$$

Si bien las entradas $\tilde{x}_i$ no aparecen en forma de producto interno, es posible reescribir la solución de otra forma mediante la fórmula de Woodburry. De este modo se obtiene que

$$\theta_{MCR} = \tilde{X}^\top \left(\tilde{X}\tilde{X}^\top + \rho \mathbb{I}\right)^{-1} Y$$

Ahora $\tilde{X}\tilde{X}^\top$ sí corresponde a un producto externo de las entradas:

$$(\tilde{X}\tilde{X}^\top)_{ij} = \langle \tilde{X}_{i\cdot}, \tilde{X}_{\cdot j}^\top \rangle = \langle \tilde{X}_{i\cdot}, \tilde{X}_{j\cdot} \rangle = \langle \tilde{x}_i, \tilde{x}_j \rangle$$

Kernel ridge regression

Otro método en el que se puede utilizar el truco del kernel es el método de regulación cuadrática usada en regresión lineal, el cual tiene solución

$$\theta_{MCR} = \left(\tilde{X}^\top \tilde{X} + \rho \mathbb{I}\right)^{-1} \tilde{X}^\top Y$$

Si bien las entradas $\tilde{x}_i$ no aparecen en forma de producto interno, es posible reescribir la solución de otra forma mediante la fórmula de Woodburry. De este modo se obtiene que

$$\theta_{MCR} = \tilde{X}^\top \left(\tilde{X}\tilde{X}^\top + \rho \mathbb{I}\right)^{-1} Y$$

Ahora $\tilde{X}\tilde{X}^\top$ sí corresponde a un producto externo de las entradas:

$$(\tilde{X}\tilde{X}^\top)_{ij} = \langle \tilde{X}_{i\cdot}, \tilde{X}_{\cdot j}^\top \rangle = \langle \tilde{X}_{i\cdot}, \tilde{X}_{j\cdot} \rangle = \langle \tilde{x}_i, \tilde{x}_j \rangle$$

Además, para un nuevo input $x_\star$, su predicción está dada por

$$\hat{y}_\star = \theta_{MCR}^\top x_\star = Y^\top \left(\tilde{X}\tilde{X}^\top + \rho \mathbb{I}\right)^{-1} \tilde{X}\tilde{x}_\star$$

donde $(\tilde{X}x_\star)_i = \langle \tilde{x}_i, x_\star \rangle$, lo cual muestra que las entradas solo aparecen en la predicción en forma de productos internos.

Kernel ridge regression

Dado un mapa de características ϕ, se denotan las características por $\phi_i = \phi(x_i)$ y $\phi_\star = \phi(x_\star)$. De este modo, se puede hacer la regresión sobre las características:

$$\Theta = \begin{pmatrix} \phi_1^\top \\ \vdots \\ \phi_N^\top \end{pmatrix} \implies \hat{y}_\star = Y^\top \left(\Theta\Theta^\top + \rho\mathbb{I} \right)^{-1} \Theta\phi_\star$$

Kernel ridge regression

Dado un mapa de características ϕ, se denotan las características por $\phi_i = \phi(x_i)$ y $\phi_\star = \phi(x_\star)$. De este modo, se puede hacer la regresión sobre las características:

$$\Theta = \begin{pmatrix} \phi_1^\top \\ \vdots \\ \phi_N^\top \end{pmatrix} \implies \hat{y}_\star = Y^\top \left(\Theta\Theta^\top + \rho\mathbb{I}\right)^{-1} \Theta\phi_\star$$

Luego, si K es un kernel asociado a ϕ: $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, se tiene que

$$\hat{y}_\star = Y^\top \left(\Theta\Theta^\top + \rho\mathbb{I}\right)^{-1} \Theta\phi_\star = Y^\top \left(K(\tilde{X}, \tilde{X}) + \rho\mathbb{I}\right)^{-1} K(\tilde{X}, x_\star),$$

donde se ha hecho abuso de notación al usar argumentos matriciales en el kernel:

$$K(\tilde{X}, \tilde{X})_{ij} = (\Theta\Theta^\top)_{ij} = \langle \phi_i, \phi_j \rangle = K(x_i, x_j)$$
$$K(\tilde{X}, x_\star)_i = (\Theta\phi_\star)_i = \langle \phi_i, \phi_\star \rangle = K(x_i, x_\star)$$