SVM con Kernel

Héctor Maravillo

Producto interno

Sea H un espacio vectorial sobre el campo $\mathbb{R}$. El **producto escalar**, **producto interno** o **producto punto** sobre H es una aplicación

$$(\cdot,\cdot) : H \times H \mapsto \mathbb{R}$$

que cumple, para todos los vectores $x, y \in H$ y cualquier escalar $\alpha \in \mathbb{R}$, las siguientes propiedades:

$$\begin{aligned}
(x, y) &= (y, x) \\
(x, y + z) &= (x, y) + (x, z) \\
(\alpha x, y) &= \alpha(x, y) \\
(x, x) &\geq 0 \\
(x, x) &= 0 \Leftrightarrow x = 0
\end{aligned}$$

Espacio pre-hilbertiano

A los espacios con **producto interno** se les conoce como **espacios pre-hilbertianos**. Son una generalización de los espacios euclidianos y su estructura permite **medir ángulos y longitudes** en el espacio vectorial.

Norma cuadrática

Sea H un espacio vectorial dotado de un producto punto. Se denomina **norma** (**norma cuadrática**) de un número $x \in H$ al número real no negativo:

$$||x|| = \sqrt{(x, x)}$$

Sea H un espacio vectorial dotado de una norma $||x||$. La función:

$$d : (x, y) \in H \times H \mapsto \mathbb{R} \tag{1}$$
$$d(x, y) := ||x - y|| \tag{2}$$

es una **distancia** sobre H.

Espacios métricos

A la tupla (X, d) donde X es un conjunto de objetos (puntos) y $d : X \times X \mapsto \mathbb{R}$ es una función que cumple que:

$$\begin{aligned} d(x,y) &\geq 0 \\ d(x,y) &= 0 \Leftrightarrow x = y \\ d(x,y) &= d(y,x) \\ d(x,y) &\leq d(x,z) + d(z,y) \end{aligned}$$

(denominada **métrica** o **distancia**) se le denomina **espacio métrico**.

Todo espacio X con una norma da lugar a un espacio métrico (X, d) definiendo:

$$d(x,y) := ||x - y||, \qquad x, y \in X$$

Convergencia en espacios métricos

Sea (X, d) un espacio métrico y $(x_n) \subset X$ una sucesión.

- La sucesión es **convergente** si existe un elemento $x_0 \in X$ tal que: $d(x_n, x_0) \to 0$ cuando $n \to \infty$. Es decir:

$$\lim_{n \to \infty} x_n = x_0$$

- La sucesión es **de Cauchy** cuando, dado $\epsilon > 0$ existe $n_0 \in \mathbb{N}$ tal que $d(x_m, x_n) < \epsilon$ para todo $m, n \geq n_0$.

Toda sucesión convergente es de Cauchy, pero en ciertos espacios no toda sucesión de Cauchy es sucesión convergente.

Espacios de Hilbert

Un espacio métrico (X, d) se dice que es **completo** si toda sucesión de Cauchy es convergente.

Se dice que un espacio con producto interno $(X, (\cdot, \cdot))$ es un **espacio de Hilbert** si el espacio métrico (X, d) es completo, con

$$\begin{aligned} ||x|| &:= \sqrt{(x, x)} \\ d(x, y) &:= ||x - y||, \qquad x, y \in X \end{aligned}$$

Espacios de Hilbert

Ejemplos de **espacios de Hilbert**:

- Espacios euclidianos $\mathbb{R}^n$ con punto interno

$$(x, y) := \sum_{i=1}^{n} x_i y_i$$

.

Espacios de Hilbert

- El espacio de sucesiones

$$l_2 := \left\{ x = (x_k) \in \mathbb{R}^n \ : \ \sum_{k=1}^{\infty} |x_k|^2 < \infty \right\}$$

con producto interno:

$$(x, y) := \sum_{i=1}^{\infty} x_i y_i$$

.

Espacios de Hilbert

- Espacios de funciones L^2 (funciones definidas en $[a, b]$ tal que la integral de $|f(x)|^2$ es finita) con producto interno:

$$(f, g) := \int_a^b f(x)g(x)dx$$

Teorema de Mercer

Theorem

Sea $x \in \mathbb{R}^l$ y Φ un mapeo $x \mapsto \Phi(x) \in H$ donde H es un espacio de Hilbert. El **producto interno** *en H tiene una* **representación equivalente**, *es decir:*

$$(\Phi(x), \Phi(z)) = K(x, z), \qquad x, z \in \mathbb{R}^l \tag{3}$$

donde $K(x, z)$ es una función **simétrica** *y* **continua** *que satisface la siguiente condición:*

$$\int_C \int_C K(x, z) g(x) g(z) dx dz \geq 0 \tag{4}$$

para cualquier función $g(x)$, $x \in C \subset \mathbb{R}^l$ tal que

$$\int_C g(x)^2 dx < \infty$$

donde C es un subconjunto compacto de $\mathbb{R}^l$

Teorema de Mercer

La afirmación inversa al teorema anterior también siempre es cierta, esto es: Para cualquier función simétrica y continua $K(x, z)$ que satisface (3) y (4) **existe** un espacio en el cual $K(x, z)$ define un producto interno.

Las funciones que cumplen con esas condiciones son conocidas como **Kernels** y el espacio H como **Espacios de Hilbert con núcleo reproductor** (*reproduccing kernel Hilbert space RKHS*).

Kernels

El teorema de Mercer **no dice como encontrar** esos espacios, solo **asegura que existen**.

No existe una herramienta general para construir el mapeo $\Phi()$ una vez que conocemos el producto interno de su espacio correspondiente.

Tampoco podemos conocer, en general, la dimensión del espacio H incluso, la cual podría ser infinita.

Construcción de Kernels

Una **condición necesaria y suficiente** para que una función $K(x, z)$ sea un kernel válido es que la **matriz de Gramm K**, cuyos elementos son $k(x_i, x_j)$ con $x_i, x_j \in X$ sea una **matriz semidefinida positiva**, es decir que:

$$v^T \mathbf{K} v \geq 0, \qquad v \in \mathbb{R}^l$$

Otra técnica poderosa es construir nuevos Kernels a partir de Kernels simples conocidos, usando las siguientes reglas.

Construcción de Kernels

Sea $K(x, z)$ un kernel, entonces las siguientes funciones también son kernels:

$$\begin{aligned} K_1(x,z) &= cK(x,z) \\ K_1(x,z) &= f(x)K(x,z)f(z) \\ K_1(x,z) &= q(K(x,z)) \\ K_1(x,z) &= \exp\left(K(x,z)\right) \end{aligned}$$

donde $c > 0$ es una constante, f es cualquier función, q es un polinomio con coeficientes no negativos.

Construcción de Kernels

Sea $K_1(x,z)$ y $K_2(x,z)$ dos kernels, entonces, las siguientes funciones también son kernels:

$$\begin{aligned} K_3(x,z) &= K_1(x,z) + K_2(x,z) \\ K_3(x,z) &= K_1(x,z)K_2(x,z) \end{aligned}$$

Kernels principales

Los ejemplos clásicos de Kernels usados en reconocimiento de patrones son:

- **Lineal**
$$K(x,z) = x^T z$$
- **Polinomiales**
$$K(x,z) = (x^T z + c)^q$$
- **Funciones de base radial (RBF) o Kernel Gaussiano**
$$K(x,z) = \exp\left(-\frac{||x-z||^2}{\sigma^2}\right)$$
- **Tangente hiperbólica** o **sigmoidal**
$$K(x,z) = \tanh\left(\gamma x^T z + c\right)$$
para valores apropiados de γ y c que satisfagan las condiciones del teorema de Mercer. Una posibilidad es usar $\gamma = 2$ y $c = 1$.

Ejemplo

Sea

$$K(x,z) = (x^T z + 1)^2$$

con $x, z \in \mathbb{R}^2$ entonces

$$\begin{aligned} k(x,z) &= (x_1 y_1 + x_2 y_2 + 1)^2 \\ &= x_1^2 y_1^2 + x_2^2 y_2^2 + 1 + 2x_1 y_1 x_2 y_2 + 2x_1 y_1 + 2x_2 y_2 \end{aligned}$$

Esto corresponde al producto interno de

$$\Phi(x) = (1,\ \sqrt{2}x_1,\ \sqrt{2}x_2, \sqrt{2}x_1 x_2,\ x_1^2,\ x_2^2)^T$$

Truco del Kernel

Para utilizar el truco del Kernel hay que definir una función de Kernel válida.

- Un enfoque es elegir un mapeo $\Phi(x)$ del espacio de atributos y usar su producto interno para encontrar un Kernel, es decir, definir

$$K(x, z) = \Phi(x)^T \Phi(z)$$

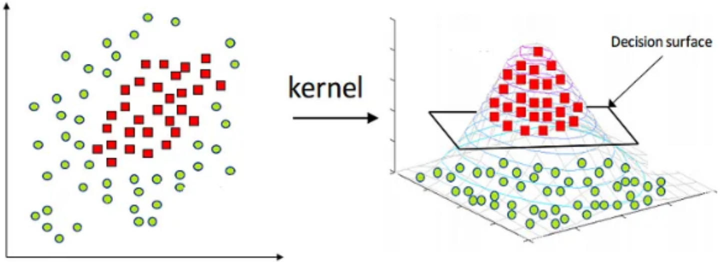
- Otro enfoque, es construir una función de Kernel directamente, sin preocuparse de identificar cual es el espacio H correspondiente.

Truco del Kernel

El concepto de un Kernel correspondiente al producto interno en un espacio de características permite **construir extensiones de algoritmos lineales**, haciendo uso del **truco del kernel** o **sustitución Kernel**.

La idea general es: si tenemos un algoritmo formulado de forma que el vector de entradas x participa sólo en forma de productos internos, podemos **remplazar el producto interno** por una función de Kernel.

Truco del Kernel



SVM con kernel

Una vez elegido un kernel apropiado, este define un mapeo en un espacio de mayor dimensión (un RKHS), y el modelo de optimización dual de Wolfe correspondiente se convierte en:

$$
\begin{aligned}
\max_{\boldsymbol{\lambda}} \quad & \left(\sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right) \\
\sum_{i=1}^{n} \lambda_i y_i &= 0 \\
0 \leq \lambda_i &\leq C \qquad i = 1, \ldots, n
\end{aligned}
$$

SVM con kernel

A partir de los multiplicadores de Lagrange óptimos encontrados en el problema anterior, el clasificador lineal en el espacio RKHS consiste en:

- Asignar x en la clase ω_1 si

$$g(x) = \sum_{i=1}^{N_s} \lambda_i y_i K(x_i, x) + b > 0$$

- Asignar a ω_2 en caso contrario.

Debido a la no linealidad de la función de Kernel, el clasificador resultante es no lineal en el espacio $\mathbb{R}^l$ original.

- La **complejidad computacional** es **independiente** de la **dimensión del espacio del Kernel**, donde el espacio de atributos es mapeado, debido al truco del Kernel.
- Uno resuelve un SVM en un espacio de alta dimensión sin tener que adoptar explícitamente el modelo en ese espacio con un gran número de parámetros.
- Por estas cuestiones, el SVM tiende a mostrar un buen rendimiento.

SVM con Kernel

Limitaciones

- La principal limitación es que no existe ningún método práctico eficiente para seleccionar la mejor función de Kernel.
- Una vez seleccionada la función de Kernel, es necesario definir sus parámetros, así como el parámetro de suavizado C.