Clase 18: Aprendizaje Estadístico

Responsable: Francisco Velasco Medina

EST-25134, Primavera 2021
Dr. Alfredo Garbuno Iñigo
23 marzo 2021

1. Generalización de Máquinas de Soporte Vectorial

Definición 1.1. Definición

Un **espacio de Hilbert** es un espacio vectorial con producto interior y completo.

Llevamos los datos de $\psi(x) = (x, x^2)$ a $signo(\langle w, \psi() \rangle - b)$.

1. Consideramos un conjunto $\mathcal{X}$ y una tarea de aprendizaje (clasificación). Usemos una función $\psi : \mathcal{X} \to \mathcal{F}$. $\mathcal{F} = \mathbb{R}^n$ puede ser un espacio de Hilbert.
2. Consideramos una muestra $S = \{(x_i, y_i)\}_{i=1}^m \to \hat{S} = \{(\psi(x_i), y_i)\}_{i=1}^m$.
3. Entrenamos con $\hat{S}$.
4. Predecimos con $(h \cdot \psi)(x) = h(\psi(x)).\psi$: enriquecer nuestra representación con los atributos.

2. El truco del kernel

Dada $\psi : \mathcal{X} \to \mathcal{F}$. Definimos el kernel como

$$\begin{aligned} k(x, x') &= \langle \psi(x), \psi(x') \rangle_{\mathcal{F}} \\ &= \langle x, x' \rangle_{\psi} \cdot \min_w L_S^H(w) + \lambda ||w||_2^2 \\ &\Rightarrow \underbrace{\min_w f(\langle w, \psi(x_1) \rangle, \langle w, \psi(x_2) \rangle + R(||w||_2)}_{\text{(1) Monótona no decreciente}}, \end{aligned} \tag{1}$$

Teorema 2.1. *Teorema del Representante*

Si $\psi : \mathcal{X} \to \mathcal{F}$ donde $\mathcal{F}$ es un espacio de Hilbert. Entonces $\exists\ \alpha \in \mathbb{R}^m$ tal que $w = \sum_{i=1}^m \alpha_i \psi(x_i)$ y además es una solución a (1).

Demostración. Sea w^* una solución a (1) y $w^* \in \mathcal{F}$

$$w^* = \sum_{i=1}^m \alpha_i \psi(x_i) + u$$

donde $\langle u, \psi(x_i) \rangle = 0, \forall\ i \in \{1, \dots, m\}$. Sea $w = w^* - u$:

$$\begin{aligned} &\Rightarrow ||w^*||_2 = ||w||_2 + ||u||_2 \\ &\Rightarrow R(||w||_2) \leq R(||w^*||_2) \\ \langle w, \psi(x_i) \rangle = \langle w^* - u, \psi(x_i) \rangle &= \langle w^*, \psi(x_i) \rangle \qquad \forall\ i \in \{1, \dots, m\} \\ &\Rightarrow f(w) = f(w^* \end{aligned}$$

$\Rightarrow w$ es óptimo y además $w = \sum_{i=1}^m \alpha_i \psi(x_i)$ □

Entonces optimizamos (1) en términos de α.

$$\begin{aligned}
\langle w, \psi(x_{i]})\rangle &= \sum_{j=1}^{m} \alpha_j \langle \psi(x_j, \psi(x_i)\rangle \\
||w||_2^2 &= \langle w, w\rangle_{\mathcal{F}} \\
&= \sum_{j,i}^{m} \alpha_j \alpha_i \langle \psi(x_j, \psi(x_i)\rangle \\
&= \sum_{j,i}^{m} \alpha_j \alpha_i k(x_j, x_i) \\
&= \alpha^T G \alpha && donde G_{ij} = k(x_j, x_i)
\end{aligned}$$

$$\Rightarrow \min_{\alpha} f(\sum \alpha_j k(x_j, x_1), \ldots, \sum \alpha_i k(x_j, x_m) + R(\sqrt{\alpha^T G \alpha})$$
$$\Rightarrow \min_{\alpha} \lambda \alpha^T G \alpha + \frac{1}{m} \sum_{i=1}^{m} \max\{0, 1 - y_i (G\alpha)_i\}$$

Ejemplo 2.2. Ejemplos de kernel Polinomios de grado k: $k(x, x') = (1 + \langle x, x'\rangle)^k; \psi(x) = (1, x, \ldots, x^k)$.

Kernel gaussiano: $k(x, x') = exp(\frac{||x-x'||^2}{-2\sigma^2}; \psi(x) = (\ldots, \frac{1}{\sqrt{n!\sigma^2}} exp\left(\frac{-x^2}{2\sigma^2}\right) x^n, \ldots$

Hiperbólico: $k(x, x') = tanh(\sigma \langle x, x'\rangle + b)$

$k(x, x') = exp\left(\frac{||x-x'||_\Sigma^2}{-2}\right); ||r||_\Sigma^2 = r^T \Sigma^{-1} r, \Sigma = diag(\sigma_1^2, \ldots, \sigma_k^2)$

$k(A, A') = 2^{|A \bigcap A'|}$

Lema 2.3. *El kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ es válido para definir un producto interior en un espacio de Hilbert. $\iff$ Es una función positiva definida. Es decir, $\forall x_1, \ldots, x_m$ la matriz ψ es una matriz positiva definida.*

2.1. Propiedades

Si k_1 y k_2 son dos kernel válidos

1. $k = k_1 + k_2$
2. $k = f(x) k_1(x, x') f(x')$
3. $k = exp(k_1(x, x'))$
4. $k = k_1 \times k_2$
5. $k = k_1(x_a, x_a') \underset{\times}{+} k_2(x_b, x_b'), x = (x_a, x_b)$