Support Vector Machines

Peng Lingwei

August 4, 2019

Contents

16 Kernel Methods

16.1 Little about Kernel Methods

Definition 1. ***(Kernels).*** *A kernel function* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

We want $K(x, x') = \langle \phi(x), \phi(x') \rangle$, where $\phi : \mathcal{X} \to \mathbb{H}$ maps $\mathcal{X}$ to Hibert space $\mathbb{H}$ called a **feature space**.

Definition 2. ***(Positive definite symmetric kernels).*** $\forall \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$, *the matrix* $\boldsymbol{K} = [K(x_i, x_j)]_{ij}$ *is symmetric positive semidefinite (SPSD).*

Example 1. *Some kernels:*

1. *Polynomial kernels:* $\forall \vec{x}, \vec{x}' \in \mathbb{R}^N, K(\vec{x}, \vec{x}') = (\langle \vec{x}, \vec{x}' \rangle + c)^d$.
2. *Gaussian kernels (Radial Basis Function, RBF):* $\forall \vec{x}, \vec{x}' \in \mathbb{R}^N, K(\vec{x}, \vec{x}') = \exp\left(-\frac{\|\vec{x}' - \vec{x}\|^2}{2\sigma^2}\right)$.
3. *Sigmoid kernels:* $\forall \vec{x}, \vec{x}' \in \mathbb{R}^N, K(\vec{x}, \vec{x}') = \tanh\left(a \langle \vec{x}, \vec{x}' \rangle + b\right)$

Lemma 1. ***(Cauchy-Schwarz inequality for PDS kernels).***

$$K(\vec{x}, \vec{x}')^2 \leq K(\vec{x}, \vec{x}) K(\vec{x}', \vec{x}')$$

Theorem 1. ***(Reproducing kernel Hibert space (RKHS)).*** *If K is a PDS kernel, then there exists a Hilbert space $\mathbb{H}$ and a mapping ϕ such that:*

$$\forall \vec{x}, \vec{x}' \in \mathcal{X}, \quad K(\vec{x}, \vec{x}') = \langle \phi(\vec{x}), \phi(\vec{x}') \rangle$$

Proof. First, we denote $\Phi_{\vec{w}}(\vec{x}) = K(\vec{w}, \vec{x})$. If the theorem is true, then we have $\Phi_{\vec{w}}(\vec{x}) = \langle \phi(\vec{w}), \phi(\vec{x}) \rangle$.
we also define subspace $\mathbb{H}_W \subset \mathbb{H}$:

$$\mathbb{H}_W = \left\{ \sum_{i \in [|W|]} a_i \Phi_{w_i} : a_i \in \mathbb{R}, w_i \in W, i \in [|W|] \right\}$$

Then, we define the inner product operation $\langle \cdot, \cdot \rangle$ on $\mathbb{H}_W \times \mathbb{H}_W$ defined for all $f, g \in \mathbb{H}_W$ with $f = \sum_{i \in I} a_i \Phi_{w_i}$ and $g = \sum_{j \in J} b_j \Phi_{w_j}$ by

$$\langle f, g \rangle = \sum_{i \in I, j \in J} a_i b_j K(w_i, w_j) = \sum_{j \in J} b_j f(w_j) = \sum_{i \in I} a_i g(w_i)$$

So

$$\langle f, f \rangle = \sum_{i,j \in I} a_i a_j K(x_i, x_j) \geq 0.$$

Then

$$\sum_{i,j=1}^{m} c_i c_j \langle f_i, f_j \rangle = \langle \sum_{i=1}^{m} c_i f_i, \sum_{j=1}^{m} c_j f_j \rangle \geq 0$$

$\square$

Definition 3. ***(Normalized kernel K).***

$$\forall \vec{x}, \vec{x}' \in \mathcal{X}, K^{norm}(\vec{x}, \vec{x}') = \frac{K(x, x')}{\sqrt{K(x,x)K(x',x')}}$$

The Gaussian kernel comes from normalizing the kernel $K = \exp\left(\frac{\langle x, x' \rangle}{\sigma^2}\right)$*.*

16.2 THE KERNEL TRICK

Definition 4. ***(General problem).*** *General problem:*

$$\min_{\vec{w}}(L(\langle \vec{w}, \vec{x}_1 \rangle, \ldots, \langle \vec{w}, \vec{x}_m \rangle)) + R(\|\vec{w}\|)$$

where $L : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$*, and R is non-decreasing function.j*

Theorem 2. *The optimal solution of general problem* $\vec{w}^* = \sum_{i=1}^m \alpha_i \phi(\vec{x}_i)$*.*

Then, the general problem can be rewritten into

$$\min_{\vec{\alpha} \in \mathbb{R}^m} L\left(\sum_{i=1}^m \alpha_i K(\vec{x}_i, \vec{x}_1), \ldots, \sum_{i=1}^m \alpha_i K(\vec{x}_i, \vec{x}_m)\right) + R\left(\sqrt{\sum_{i,j=1}^m \alpha_i \alpha_j K(\vec{x}_i, \vec{x}_j)}\right)$$

Let $\mathbf{K}_{ij} = K(\vec{x}_i, \vec{x}_j)$, then the Soft-SVM can be rewritten into

$$\min_{\vec{\alpha} \in \mathbb{R}^m} \left(\lambda \vec{\alpha}^T \mathbf{K} \vec{\alpha} + \frac{1}{m} \sum_{i=1}^m \max\{0, 1 - y_i(\mathbf{K}\vec{\alpha})_i\}\right)$$

We can calculate the prediction by

$$h_{\vec{w}}(\vec{x}) = \langle \vec{w}, \phi(\vec{x}) \rangle = \sum_{j=1}^m \alpha_i \langle \phi(\vec{x}_i), \vec{x} \rangle = \sum_{i=1}^m \alpha_i K(\vec{x}_i, \vec{x})$$