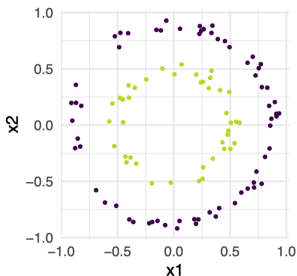


Introduction to Machine Learning

Reproducing Kernel Hilbert Space and Representer Theorem



Learning goals

- Know that for every kernel there is an associated feature map and space (Mercer's Theorem)
- Know that this feature map is not unique, and the reproducing kernel Hilbert space (RKHS) is a reference space
- Know the representation of the solution of a SVM is given by the representer theorem

KERNELS: MERCER'S THEOREM

- Kernels are symmetric, positive definite functions $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.
- A kernel can be thought of as a shortcut computation for a two-step procedure: the feature map and the inner product.

Mercer's theorem says that for every kernel there exists an associated (well-behaved) feature space where the kernel acts as a dot-product.

- There exists a Hilbert space Φ of continuous functions $\mathcal{X} \rightarrow \mathbb{R}$ (think of it as a vector space with inner product where all operations are meaningful, including taking limits of sequences; this is non-trivial in the infinite-dimensional case)
- and a continuous “feature map” $\phi : \mathcal{X} \rightarrow \Phi$,
- so that the kernel computes the inner product of the features:

$$k(\mathbf{x}, \tilde{\mathbf{x}}) = \langle \phi(\mathbf{x}), \phi(\tilde{\mathbf{x}}) \rangle .$$

REPRODUCING KERNEL HILBERT SPACE

- There are many possible Hilbert spaces and feature maps for the same kernel, but they are all “equivalent” (isomorphic).
- It is often helpful to have a reference space for a kernel $k(\cdot, \cdot)$, called the **reproducing kernel Hilbert space (RKHS)**.
- The feature map of this space is

$$\phi : \mathcal{X} \rightarrow \mathcal{C}(\mathcal{X}); \quad \mathbf{x} \mapsto k(\mathbf{x}, \cdot) ,$$

where $\mathcal{C}(\mathcal{X})$ is the space of continuous functions $\mathcal{X} \rightarrow \mathbb{R}$. The "features" of the RKHS are the kernel functions evaluated at an \mathbf{x} .

- The Hilbert space is the completion of the span of the features:

$$\Phi = \overline{\text{span}\{\phi(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}} \subset \mathcal{C}(\mathcal{X}) .$$

- The so-called **reproducing property** states:

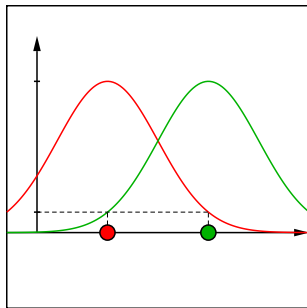
$$\langle k(\mathbf{x}, \cdot), k(\tilde{\mathbf{x}}, \cdot) \rangle = \langle \phi(\mathbf{x}), \phi(\tilde{\mathbf{x}}) \rangle = k(\mathbf{x}, \tilde{\mathbf{x}}).$$

REPRODUCING KERNEL HILBERT SPACE

- The RKHS provides us with a useful interpretation:
an input $\mathbf{x} \in \mathcal{X}$ mapped to the **basis function** $\phi(\mathbf{x}) = k(\mathbf{x}, \cdot)$.
- The kernel maps 2 points and computes the inner product:

$$\langle k(\mathbf{x}, \cdot), k(\tilde{\mathbf{x}}, \cdot) \rangle = k(\mathbf{x}, \tilde{\mathbf{x}}) .$$

- This is best illustrated with the Gaussian kernel.



REPRODUCING KERNEL HILBERT SPACE

- Caveat: Not all elements of the Hilbert space are of the form $k(\mathbf{x}, \cdot)$ for some $\mathbf{x} \in \mathcal{X}$!
- A general element in the span takes the form

$$\sum_{i=1}^n \alpha_i k(\mathbf{x}^{(i)}, \cdot) \in \Phi .$$

- A general element in the closure of the span takes the form

$$\sum_{i=1}^{\infty} \alpha_i k(\mathbf{x}^{(i)}, \cdot) \in \Phi .$$

with $\sum_{i=1}^{\infty} \alpha_i^2 < \infty$.

REPRODUCING KERNEL HILBERT SPACE

What is $\langle f, g \rangle$ for two elements

$$f = \sum_{i=1}^n \alpha_i k(\mathbf{x}^{(i)}, \cdot), \quad g = \sum_{j=1}^m \beta_j k(\mathbf{x}^{(j)}, \cdot) ?$$

We use the bilinearity of the inner product:

$$\begin{aligned} \left\langle \sum_{i=1}^n \alpha_i k(\mathbf{x}^{(i)}, \cdot), \sum_{j=1}^m \beta_j k(\mathbf{x}^{(j)}, \cdot) \right\rangle &= \sum_{i=1}^n \alpha_i \left\langle k(\mathbf{x}^{(i)}, \cdot), \sum_{j=1}^m \beta_j k(\mathbf{x}^{(j)}, \cdot) \right\rangle \\ &= \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j \langle k(\mathbf{x}^{(i)}, \cdot), k(\mathbf{x}^{(j)}, \cdot) \rangle \\ &= \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \end{aligned}$$

The kernel defines the inner products of all elements in the span of the basis functions.

REPRESENTER THEOREM

The **representer theorem** tells us that the solution of a support vector machine problem

$$\begin{aligned} \min_{\boldsymbol{\theta}, \theta_0, \zeta^{(i)}} \quad & \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + c \sum_{i=1}^n \zeta^{(i)} \\ \text{s.t.} \quad & y^{(i)} \left(\left\langle \boldsymbol{\theta}, \phi \left(\mathbf{x}^{(i)} \right) \right\rangle + \theta_0 \right) \geq 1 - \zeta^{(i)} \quad \forall i \in \{1, \dots, n\}, \\ \text{and} \quad & \zeta^{(i)} \geq 0 \quad \forall i \in \{1, \dots, n\} \end{aligned}$$

can be written as

$$\boldsymbol{\theta} = \sum_{j=1}^n \beta_j \phi \left(\mathbf{x}^{(j)} \right)$$

for $\beta_j \in \mathbb{R}$.

REPRESENTER THEOREM

Theorem (Representer Theorem):

The solution θ, θ_0 of the support vector machine optimization problem fulfills $\theta \in V = \text{span} \{ \phi(\mathbf{x}^{(1)}) , \dots , \phi(\mathbf{x}^{(n)}) \}$.

Proof: Let V^\perp denote the space orthogonal to V , so that $\Phi = V \oplus V^\perp$. The vector θ has a unique decomposition into components $\mathbf{v} \in V$ and $\mathbf{v}^\perp \in V^\perp$, so that $\mathbf{v} + \mathbf{v}^\perp = \theta$.

The regularizer becomes $\|\theta\|^2 = \|\mathbf{v}\|^2 + \|\mathbf{v}^\perp\|^2$. The constraints $y^{(i)} \left(\langle \theta, \phi(\mathbf{x}^{(i)}) \rangle + \theta_0 \right) \geq 1 - \zeta^{(i)}$ do not depend on \mathbf{v}^\perp at all:

$$\langle \theta, \phi(\mathbf{x}^{(i)}) \rangle = \langle \mathbf{v}, \phi(\mathbf{x}^{(i)}) \rangle + \underbrace{\langle \mathbf{v}^\perp, \phi(\mathbf{x}^{(i)}) \rangle}_{=0} \quad \forall i \in \{1, 2, \dots, n\}.$$

Thus, we have two independent optimization problems, namely the standard SVM problem for \mathbf{v} and the unconstrained minimization problem of $\|\mathbf{v}^\perp\|^2$ for \mathbf{v}^\perp , with obvious solution $\mathbf{v}^\perp = 0$. Thus, $\theta = \mathbf{v} \in V$.

REPRESENTER THEOREM

- Hence, we can restrict the SVM optimization problem to the **finite-dimensional** subspace span $\{ \phi(\mathbf{x}^{(1)}) , \dots , \phi(\mathbf{x}^{(n)}) \}$. Its dimension grows with the size of the training set.
- More explicitly, we can assume the form

$$\boldsymbol{\theta} = \sum_{j=1}^n \beta_j \cdot \phi(\mathbf{x}^{(j)})$$

for the weight vector $\boldsymbol{\theta} \in \Phi$.

- The SVM prediction on $\mathbf{x} \in \mathcal{X}$ can be computed as

$$f(\mathbf{x}) = \sum_{j=1}^n \beta_j \langle \phi(\mathbf{x}^{(j)}) , \phi(\mathbf{x}) \rangle + \theta_0 .$$

It can be shown that the sum is **sparse**: $\beta_j = 0$ for non-support vectors.