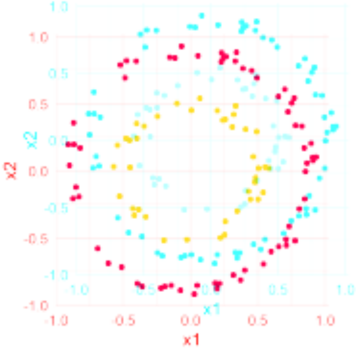
Introduction to Machine Learning

Nonlinear Support Vector Machines

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

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

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

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

where $\mathcal{C}(\mathcal{X})$ is the space of continuous functions $\mathcal{X} \to \mathbb{R}$. 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:

$$\Phi = \overline{\mathrm{span}\{\phi(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{X}\}} \subset \mathcal{C}(\mathcal{X}) \ .$$

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

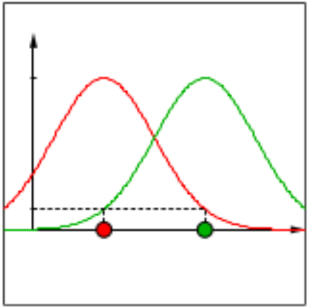
$$\langle k(\mathbf{x},\cdot), k(\tilde{\mathbf{x}},\cdot)\rangle = \langle \phi(\mathbf{x}), \phi(\tilde{\mathbf{x}})\rangle = k(\mathbf{x},\tilde{\mathbf{x}}).$$

REPRODUCING KERNEL HILBERT SPACE / 2

- 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 / 3

- 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

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

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

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

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

REPRODUCING KERNEL HILBERT SPACE / 4

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

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

We use the bilinearity of the inner product:

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

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

REPRESENTER THEOREM / 2

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

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

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

for the weight vector $\theta \in \Phi$.

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

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

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

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