Support Vector Machines - IV

Prof. Dan A. Simovici

UMB

1. Hilbert Spaces

2. Kernels

3. Positive Definite Symmetric (PDS) Kernels

4. Multiclass Classification and SVMs

What is a Hilbert Space?

Hilbert spaces are generalizations of Euclidean spaces.
A Hilbert space is a linear space that is equipped with an inner product such that the metric space generated by the inner product is complete.
The inner product of two elements x, y of a Hilbert space H is denoted by (x, y). Note that in the case of $\mathbb{R}^n$ (which is a special case of a Hilbert space) the inner product of $\boldsymbol{x}, \boldsymbol{y}$ was denoted by $\boldsymbol{x}'\boldsymbol{y}$.

Definition

A kernel over $\mathcal{X}$ is a function $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ such that there exists a function $\Phi : \mathcal{X} \longrightarrow H$ that satisfies the condition

$$K(u, v) = \langle \Phi(u), \Phi(v) \rangle,$$

where H is a Hilbert space called the feature space.

Recall the general form of the dual optimization problem for SVMs:

$$\begin{aligned}&\textit{maximize for } \boldsymbol{a} \sum_{i=1}^{m} a_i - \frac{1}{2} a_i a_j y_i y_j \boldsymbol{x}_i' \boldsymbol{x}_j \\ &\textit{subject to } 0 \leqslant a_i \leqslant C \textit{ and } \sum_{i=1}^{m} a_i y_i = 0 \\ &\textit{for } 1 \leqslant i \leqslant m.\end{aligned}$$

Note the presence of the inner product $\boldsymbol{x}_i' \boldsymbol{x}_j$. This is replaced by the inner product $(\Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j))$, in the Hilbert feature space, that is, by $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where K is a suitable kernel function.

A More General SVM Formulation

$$\begin{aligned} &\textit{maximize for } \boldsymbol{a} \ \textstyle\sum_{i=1}^m a_i - \frac{1}{2} a_i a_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &\textit{subject to } 0 \leqslant a_i \leqslant C \textit{ and } \textstyle\sum_{i=1}^m a_i y_i = 0 \\ &\textit{for } 1 \leqslant i \leqslant m. \end{aligned}$$

The hypothesis returned by the SVM algorithm is now

$$h(\boldsymbol{x}) = sign\ \left(\sum_{i=1}^m a_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \right).$$

with $b = y_i - \sum_{j=1}^m a_j y_j K(x_j, x_i)$ for any $\boldsymbol{x}_i$ with $0 < a_i < C$.
Note that we do not work with the feature mapping Φ; instead we use the kernel only!

Mercer's Theorem

Theorem

Let $\mathcal{X} \subseteq \mathbb{R}^n$ be a compact set and let $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ be a continuous and symmetric function. Then K admits a uniformly convergent expression

$$K(u, v) = \sum_{n=0}^{\infty} a_n \phi_n(u) \phi_n(v)$$

with $a_n > 0$ if and only if for every square integrable function $c \in L_2(\mathcal{X})$ we have

$$\int \int_{\mathcal{X} \times \mathcal{X}} c(u) c(v) K(u, v) \, du \, dv \geqslant 0$$

This is qquivalent to saying that the kernel is positive definite symmetric (PDS).

Definition

A kernel $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is PDS if for any $\{x_1, \ldots, x_m\} \subseteq \mathcal{X}$ the matrix $\boldsymbol{K} = (K(x_i, x_j))$ is symmetric and positive semidefinite.

A symmetric matrix $\boldsymbol{K}$ is positive semidefinite if one of the equivalent conditions:

- the eigenvalues of $\boldsymbol{K}$ are non-negative, or
- for any $\boldsymbol{c} \in \mathbb{R}^m$, $\boldsymbol{c}'\boldsymbol{K}\boldsymbol{c} \geqslant 0$

hold.

Example

For $c > 0$ a polynomial kernel of degree d is the kernel defined over $\mathbb{R}^n$ by

$$K(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}'\boldsymbol{v} + c)^d.$$

As an example, consider $n = 2$, $d = 2$ and the kernel $K(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}'\boldsymbol{v} + c)^2$. We have

$$\begin{aligned} K(\boldsymbol{u}, \boldsymbol{v}) &= (u_1 v_1 + u_2 v_2 + c)^2 \\ &= u_1^2 v_1^2 + u_2^2 v_2^2 + c^2 + 2u_1 v_1 u_2 v_2 + 2u_1 v_1 c + 2u_2 v_2 c, \end{aligned}$$

Example (cont'd)

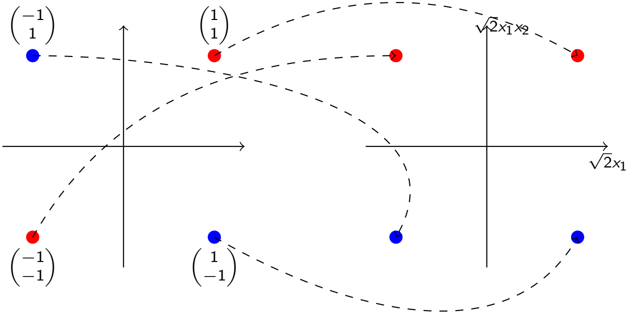
Feature space is $\mathbb{R}^6$

$$K(\boldsymbol{u},\boldsymbol{v}) = \begin{pmatrix} u_1^2 \\ u_2^2 \\ \sqrt{2}u_1 u_2 \\ \sqrt{2c}u_1 \\ \sqrt{2c}u_2 \\ c \end{pmatrix}' \begin{pmatrix} v_1^2 \\ v_2^2 \\ \sqrt{2}v_1 v_2 \\ \sqrt{2c}v_1 \\ \sqrt{2c}v_2 \\ c \end{pmatrix} = \Phi(\boldsymbol{u})'\Phi(\boldsymbol{v}) \text{ and } \Phi(\boldsymbol{x}) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1 x_2 \\ \sqrt{2c}x_1 \\ \sqrt{2c}x_2 \\ c \end{pmatrix}$$

In general, features associated to a polynomial kernel of degree d are all monomials of degree d associated to the original features. It is possible to show that polynomial kernels of degree d on $\mathbb{R}^n$ map the input space to a space of dimension $\binom{n+d}{d}$.

For the kernel $K(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}'\boldsymbol{v} + 1)^2$ we have

$$\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} x_1 x_2 \\ \sqrt{2} x_1 \\ \sqrt{2} x_2 \\ 1 \end{pmatrix}.$$


$\begin{pmatrix}-1\\1\end{pmatrix}$
$\begin{pmatrix}1\\1\end{pmatrix}$
$\begin{pmatrix}-1\\-1\end{pmatrix}$
$\begin{pmatrix}1\\-1\end{pmatrix}$
$\sqrt{2}x_1x_2$
$\sqrt{2}x_1$

For the kernel $K(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{u}'\boldsymbol{v}+1)^2$ we have

$$\Phi\begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}1\\1\\\sqrt{2}\\\sqrt{2}\\\sqrt{2}\\1\end{pmatrix}, \Phi\begin{pmatrix}-1\\-1\end{pmatrix} = \begin{pmatrix}1\\1\\\sqrt{2}\\-\sqrt{2}\\-\sqrt{2}\\1\end{pmatrix}, \Phi\begin{pmatrix}-1\\1\end{pmatrix} = \begin{pmatrix}1\\1\\-\sqrt{2}\\-\sqrt{2}\\\sqrt{2}\\1\end{pmatrix}, \Phi\begin{pmatrix}1\\-1\end{pmatrix} = \begin{pmatrix}1\\1\\-\sqrt{2}\\\sqrt{2}\\-\sqrt{2}\\1\end{pmatrix}$$

For this set of points differences occur in the third,fourth, and fifth features.

Example

For $a, b \geqslant 0$, a *sigmoid kernel* is defined as

$$K(\boldsymbol{x}, \boldsymbol{y}) = \tanh(a\boldsymbol{x}'\boldsymbol{y} + b)$$

With $a, b \geqslant 0$ the kernel is PDS.

Definition

To any kernel K we can associate a normalized kernel K' defined by

$$K'(u,v) = \begin{cases} 0 & \text{if } K(u,u) = 0 \text{ or } K(v,v) = 0, \\ \frac{K(u,v)}{\sqrt{K(u,u)}\sqrt{K(v,v)}} & \text{otherwise.} \end{cases}$$

If $K(u,u) \neq 0$, then $K'(u,u) = 1$.

Example

Let K be the kernel

$$K(\boldsymbol{u}, \boldsymbol{v}) = e^{\frac{\boldsymbol{u}'\boldsymbol{v}}{\sigma^2}},$$

where $\sigma > 0$. Note that $K(\boldsymbol{u}, \boldsymbol{u}) = e^{\frac{\|\boldsymbol{u}\|^2}{\sigma^2}}$ and $K(\boldsymbol{v}, \boldsymbol{v}) = e^{\frac{\|\boldsymbol{v}\|^2}{\sigma^2}}$, hence its normalized kernel is

$$\begin{aligned} K'(\boldsymbol{u}, \boldsymbol{v}) &= \frac{K(u, v)}{\sqrt{K(u, u)}\sqrt{K(v, v)}} \\ &= \frac{e^{\frac{\boldsymbol{u}'\boldsymbol{v}}{\sigma^2}}}{e^{\frac{\|\boldsymbol{u}\|^2}{2\sigma^2}} e^{\frac{\|\boldsymbol{v}\|^2}{2\sigma^2}}} \\ &= e^{-\frac{\|\boldsymbol{u}-\boldsymbol{v}\|^2}{2\sigma^2}} \end{aligned}$$

Example

For a positive constant σ a Gaussian kernel or a radial basis function is the function $K : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ defined by

$$K(\boldsymbol{u}, \boldsymbol{v}) = e^{-\frac{\|\boldsymbol{u}-\boldsymbol{v}\|^2}{2\sigma^2}}.$$

Theorem

Let K be a PDS kernel. For any $u, v \in \mathcal{X}$ we have

$$K(u,v)^2 \leqslant K(u,u)K(v,v).$$

Proof: Consider the matrix

$$\boldsymbol{K} = \begin{pmatrix} K(u,u) & K(u,v) \\ K(v,u) & K(v,v) \end{pmatrix}$$

$\boldsymbol{K}$ is positive semidefinite, so its eigenvalues λ_1, λ_2 must be non-negative. Its characteristic equation is

$$\begin{vmatrix} K(u,u) - \lambda & K(u,v) \\ K(v,u) & K(v,v) - \lambda \end{vmatrix} = 0$$

Equivalently,

$$\lambda^2 - (K(u,u) + K(v,v))\lambda + \det(\boldsymbol{K}) = 0$$

Therefore, $\lambda_1\lambda_2 = \det(\boldsymbol{K}) \geqslant 0$ and this implies

$$K(u,u)K(v,v) - K(u,v)^2 \leqslant 0.$$

Theorem

Let K be a PDS kernel. Its normalized kernel is PDS.

Proof: Let $\{x_1, \ldots, x_m\} \subseteq \mathcal{X}$ and $\boldsymbol{c} \in \mathbb{R}^m$. We prove that $\sum_{i,j} c_i c_j K'(x_i, x_j) \geqslant 0$.
If $K(x_i, x_i) = 0$, then $K(x_i, x_j) = 0$ and, thus, $K'(x_i, x_j) = 0$ for $1 \leqslant j \leqslant m$.
Thus, we may assume that $K(x_i, x_i) > 0$ for $1 \leqslant i \leqslant m$. We have

$$\begin{aligned}
\sum_{i,j} c_i c_j K'(x_i, x_j) &= \sum_{i,j} c_i c_j \frac{K(x_i, x_j)}{\sqrt{K(x_i, x_i) K(x_j, x_j)}} \\
&= \sum_{i,j} c_i c_j \frac{\langle \Phi(x_i), \Phi(x_j) \rangle}{\| \Phi(x_i) \|_H \| \Phi(x_j) \|_H} \\
&= \left\| \sum_i \frac{c_i \Phi(x_i)}{\| \Phi(x_i) \|_H} \right\| \geqslant 0,
\end{aligned}$$

where Φ is the feature mapping associated to K.

Theorem

Let $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ be a PDS kernel. Then, there exists a Hilbert space H of functions and a feature mapping $\Phi : \mathcal{X} \longrightarrow H$ such that $K(\boldsymbol{x}, \boldsymbol{y}) = (\Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}))$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$. Furthermore, H has the reproducing property which means that for every $h \in H$ we have

$$h(\boldsymbol{x}) = (h, K(\boldsymbol{x}, \cdot)).$$

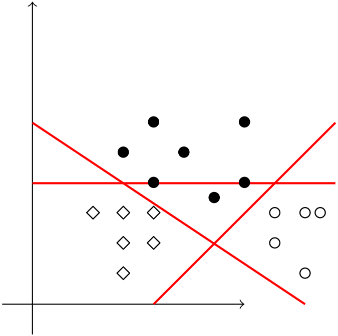
The function space H is called a reproducing Hilbert space associated with K.

In the standard approach, the two-class decision functions can be extended to k classes by constructing k decision functions for each of the k classes, where

$$h_i(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in C_i, \\ -1 & \text{otherwise} \end{cases}$$

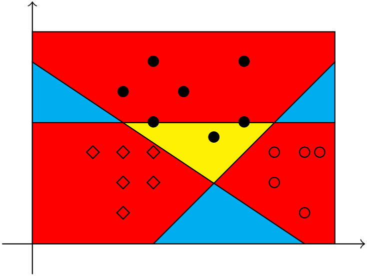
for $1 \leqslant i \leqslant k$.

Consider the use of linear classifiers for a three class problem.



The lines divide the plane in seven regions.

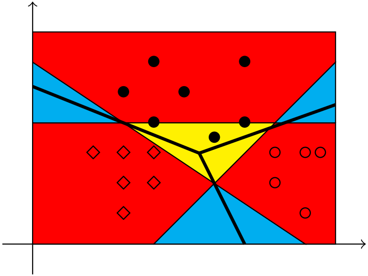
Only in the red regions exactly one decision function is 1.



In the cyan regions two decision functions equal 1, and in the yellow region no decision functions equals 1.

The *winner-takes-all* approach for several classes

To break ties (as in the cyan regions) one can drop the `sign` from the hyptheses and use the real input values to the `sign` instead.
The classification result is the index of the largest value of hypotheses



(without the `signum`).

The *pairwise* classification

For the pairwise classification a decision function h_{kl} is defined for each pair (k, l) of classes. Since the pairwise approach is symmetric we have $h_{kl} = -h_{lk}$. As a notational device we also define $f_{kk} = 0$. Thus, we have

$$f_{kl}(\boldsymbol{x}) = \begin{cases} 1 & \text{for all examples in class } C_k, \\ -1 & \text{for all examples in class } C_l. \end{cases}$$

There exist $\binom{k}{2}$ different pairwise decision functions.
The class can be calculated by summing up the decision functions

$$f_k = \sum_l f_{kl}.$$

The class k of $\boldsymbol{x}$ is given by $k = \arg\max f_k(\boldsymbol{x})$. If there are no ties, then $\max f_k = K - 1$.