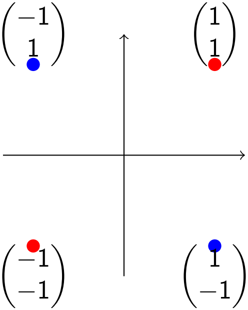
Support Vector Machines - II

Prof. Dan A. Simovici

UMB

Consider a simple data set that consists of four points in $\mathbb{R}^2$:



It is impossible to separate the red point (the positive examples) from the negative examples (the blue points) using a line, no matter how you draw the line!

Reminder: eigenvalues and eigenvectors of a matrix

Definition

An eigenvalue for a matrix $A \in \mathbb{C}^{n \times n}$ is a number λ such that

$$A\mathbf{x} = \lambda \mathbf{x}$$

for some non-zero vector $\mathbf{x} \in \mathbb{C}^n$ referred to as an *eigenvector* for λ.

This implies $\mathbf{x}^{\mathsf{H}} A\mathbf{x} = \lambda \mathbf{x}^{\mathbf{H}}\mathbf{x}$, so

$$\lambda = \frac{\mathbf{x}^{\mathsf{H}} A\mathbf{x}}{\mathbf{x}^{\mathsf{H}}\mathbf{x}}.$$

For real matrices we have

$$\lambda = \frac{\mathbf{x}' A\mathbf{x}}{\mathbf{x}'\mathbf{x}}.$$

The Characteristic Polynomial of a Matrix

If λ is an eigenvalue of the matrix $A \in \mathbb{C}^{n\times n}$, there exists a non-zero eigenvector $\mathbf{x} \in \mathbb{C}^n$ such that $A\mathbf{x} = \lambda\mathbf{x}$. Therefore, the linear system

$$(\lambda I_n - A)\mathbf{x} = \mathbf{0}_n$$

has a non-trivial solution. This is possible if and only if $\det(\lambda I_n - A) = 0$, so eigenvalues are the solutions of the equation

$$\det(\lambda I_n - A) = 0.$$

$\det(\lambda I_n - A)$ is a polynomial of degree n in λ, known as the *characteristic polynomial* matrix A. We denote this polynomial by p_A.

Example

The characteristic polynomial of the matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is

$$\begin{aligned} p(\lambda) &= \det(I_2\lambda - A) = \begin{vmatrix} \lambda - a & -b \\ -c & \lambda - d \end{vmatrix} \\ &= (\lambda - a)(\lambda - d) - bc = \lambda^2 - (a+d)\lambda + ad - bc. \end{aligned}$$

Thus, the eigenvalues are

$$\lambda_{1,2} = \frac{a + d \pm \sqrt{(a-d)^2 + 4bc}}{2}.$$

Example

Let

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

be a matrix in $\mathbb{C}^{3\times 3}$. Its characteristic polynomial is

$$\begin{aligned} p_A &= \begin{vmatrix} \lambda - a_{11} & -a_{12} & -a_{13} \\ -a_{21} & \lambda - a_{22} & -a_{23} \\ -a_{31} & -a_{32} & \lambda - a_{33} \end{vmatrix} = \lambda^3 - (a_{11} + a_{22} + a_{33})\lambda^2 \\ &+ (a_{11}a_{22} + a_{22}a_{33} + a_{33}a_{11} - a_{12}a_{21} - a_{23}a_{32} - a_{13}a_{31})\lambda \\ &- (a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{32}a_{21} - a_{12}a_{21}a_{33} - a_{23}a_{32}a_{11} - \end{aligned}$$

Recall that a matrix $A \in \mathbb{R}^{n \times n}$ is positive definite if $\mathbf{x}' A \mathbf{x} > 0$ for $\mathbf{x} \neq 0$.

Theorem

The eigenvalues of a real symmetric positive matrix are positive.

Proof: The eigenvalues of real symmetric matrices are real. If λ is an eigenvalue of A with the eigenvector $\mathbf{x}$, then $A\mathbf{x} = \lambda \mathbf{x}$, hence $\mathbf{x}' A \mathbf{x} = \lambda \mathbf{x}' \mathbf{x} = \lambda \parallel \mathbf{x} \parallel^2 > 0$. Thus, $\lambda > 0$.

Theorem

If the eigenvalues if a real symmetric matrix are positive, then A is positive definite.

Proof: For a real symmetric matrix there exists an orthogonal matrix Q such that $Q'AQ = D$, where

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

If $\mathbf{x} \neq \mathbf{0}_n$, then $\mathbf{x}'A\mathbf{x} = \mathbf{x}'Q'DQ\mathbf{x} = \mathbf{y}'D\mathbf{y}$, where $\mathbf{y} = Q\mathbf{x}$. Then, $\mathbf{y}'D\mathbf{y} = \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2 > 0$ beacuse $\mathbf{y} = Q'\mathbf{x}$ is a non-zero vector. Here we used the fact that $Q^{-1} = Q'$.

Hilbert spaces, named after David Hilbert, generalize the notion of Euclidean space. They extend the methods of vector algebra and calculus from the two-dimensional Euclidean plane and three-dimensional space to spaces with any finite or infinite number of dimensions.

- An inner product (x, y) defined on a linear space H generates a norm $\| x \| = \sqrt{(x, x)}$.
- A norm on a linear space generates a distance (a metric) $d(x, y) = \| x - y \|$. Thus, every normed space becomes a metric space.
- A *Cauchy sequence* in a metric space is a sequence (x_n) such that for every $\epsilon > 0$ there exists a number n_ϵ such that $m, p > n_\epsilon$ imply $d(x_m, x_p) < \epsilon$.
- A metric space is complete if every Cauchy sequence has a limit in that space.

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.
As above, 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 $\mathbf{x}, \mathbf{y}$ was denoted by $\mathbf{x}'\mathbf{y}$.

Example

The Euclidean space $\mathbb{R}^n$ equipped with the inner product

$$(\mathbf{x}, \mathbf{y}) = x_1 y_1 + \cdots + x_n y_n$$

is a Hilbert space.

Example

The space ℓ^2 that consists of infinite sequences of the form $\mathbf{z} = (z_1, z_2, \ldots)$ such that the series $\sum_n |z_n|^2$ converges is a Hilbert space, where the innner product is defined as

$$(\mathbf{z}, \mathbf{w}) = \sum_{n=1}^{\infty} z_n \overline{w_n}.$$

Example

For two function f, g such that $\int_a^b f^2(x)\, dx$ and $\int_a^b g^2(x)\, dx$ exist, an inner product can be defined as

$$(f, g) = \int_a^b f(x)g(x)\, dx.$$

The resulting linear space is a Hilbert space.

Definition

Let H is a Hilbert space called the feature space and let $\mathcal{X}$ be the input space that is mapped by a function $\Phi : \mathcal{X} \longrightarrow H$ into a Hilbert space.
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

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

for every $u, v \in \mathcal{X}$.

- The purpose of Φ is to map the input space $\mathcal{X}$ into a Hilbert space where data may become lineraly separable.
- If a kernel K exists, then the inner product $\langle \Phi(\mathbf{u}), \Phi(\mathbf{v}) \rangle$ in the Hilbert space that may be difficult to calculate. This is the case because we would have to compute both $\Phi(\mathbf{u})$ and $\Phi(\mathbf{v})$ and then compute the inner product $\langle \Phi(u), \Phi(v) \rangle$ in the Hilbert space. But, if there exists a kernel K, the inner product $\langle \Phi(u), \Phi(v) \rangle$ may be obtained directly using the equality $K(u, v) = \langle \Phi(u), \Phi(v) \rangle$.

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

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

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

A More General SVM Formulation

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

The hypothesis returned by the SVM algorithm is now

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

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

Definition

Let S be a non-empty set. A complex-valued function $K : S \times S \longrightarrow \mathbb{C}$ is of *positive type* if for every $n \geqslant 1$ we have:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i K(x_i, x_j) \overline{a_j} \geqslant 0$$

for every $a_i \in \mathbb{C}$ and $x_i \in S$, where $1 \leqslant i \leqslant n$.

$K : S \times S \longrightarrow \mathbb{R}$ is real and of positive type if for every $n \geqslant 1$ we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i K(x_i, x_j) a_j \geqslant 0$$

for every $a_i \in \mathbb{R}$ and $x_i \in S$, where $1 \leqslant i \leqslant n$.

If $K : S \times S \longrightarrow \mathbb{C}$ is of positive type, then taking $n = 1$ we have $aK(x,x)\overline{a} = K(x,x)|a|^2 \geqslant 0$ for every $a \in \mathbb{C}$ and $x \in S$. This implies $K(x,x) \geqslant 0$ for $x \in S$.
Note that $K : S \times S \longrightarrow \mathbb{C}$ is of positive type if for every $n \geqslant 1$ and for every $x_1, \ldots, x_s$ the matrix $A_{n,K}(x_1, \ldots, x_n) = (K(x_i, x_j))$ is positive definite, and, therefore it has positive eigenvalues.

Example

The function $K : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ given by $K(x, y) = \cos(x - y)$ is of positive type because

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{n} a_i K(x_i, x_j)\overline{a_j} &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \cos(x_i - x_j)\overline{a_j} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i(\cos x_i \cos x_j + \sin x_i \sin x_j)\overline{a_j} \\
&= \left| \sum_{i=1}^{n} a_i \cos x_i \right|^2 + \left| \sum_{i=1}^{n} a_i \sin x_i \right|^2 .
\end{aligned}$$

for every $a_i \in \mathbb{C}$ and $x_i \in S$, where $1 \leqslant i \leqslant n$.

Definition

Let S be a non-empty set. A complex-valued function $K : S \times S \longrightarrow \mathbb{C}$ is *Hermitian* if $K(x, y) = \overline{K(y, x)}$ for every $x, y \in S$.

Theorem

Let H be a Hilbert space, S be a non-empty set and let $f : S \longrightarrow H$ be a function. The function $K : S \times S \longrightarrow \mathbb{C}$ defined by

$$K(s,t) = (f(s), f(t))$$

is of positive type.

Proof

We can write

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i\overline{a_j}K(t_i,t_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i\overline{a_j}(f(t_i),f(t_j))\\ &= \left\|\sum_{i=1}^{n} a_i f(a_i)\right\|^2 \geqslant 0,\end{aligned}$$

which means that K is of positive type.

Theorem

Let S be a set and let $F : S \times S \longrightarrow \mathbb{C}$ be a positive type function. The following statements hold:

1. $F(x, y) = \overline{F(y, x)}$ *for every* $x, y \in S$, *that is,* F *is Hermitian;*
2. $\overline{F}$ *is a positive type function;*
3. $|F(x, y)|^2 \leqslant F(x, x)F(y, y)$.

Proof

Take $n = 2$ in the definition of positive type functions. We have

$$a_1\overline{a_1}F(x_1, x_1) + a_1\overline{a_2}F(x_1, x_2) + a_2\overline{a_1}F(x_2, x_1) + a_2\overline{a_2}F(x_2, x_2) \geqslant 0, \tag{1}$$

which amounts to

$$|a_1|^2F(x_1, x_1) + a_1\overline{a_2}F(x_1, x_2) + a_2\overline{a_1}F(x_2, x_1) + |a_2|^2F(x_2, x_2) \geqslant 0,$$

By taking $a_1 = a_2 = 1$ we obtain

$$p = F(x_1, x_1) + F(x_1, x_2) + F(x_2, x_1) + F(x_2, x_2) \geqslant 0,$$

where p is a positive real number.
Similarly, by taking $a_1 = i$ and $a_2 = 1$ we have

$$q = -F(x_1, x_1) + iF(x_1, x_2) - iF(x_2, x_1) + F(x_2, x_2) \geqslant 0,$$

where q is a positive real number.

Proof (cont'd)

Thus, we have

$$\begin{aligned} F(x_1, x_2) + F(x_2, x_1) &= p - F(x_1, x_1) - F(x_2, x_2), \\ iF(x_1, x_2) - iF(x_2, x_1) &= q + F(x_1, x_1) - F(x_2, x_2). \end{aligned}$$

These equalities imply

$$\begin{aligned} 2F(x_1, x_2) &= P - iQ \\ 2F(x_2, x_1) &= P + iQ, \end{aligned}$$

where $P = p - F(x_1, x_1) - F(x_2, x_2)$ and $Q = q + F(x_1, x_1) - F(x_2, x_2)$, which shows the first statement holds.

The second part of the theorem follows by applying the conjugation in the equality of Definition.
For the final part, note that if $F(x_1, x_2) = 0$ the desired inequality holds immediately. Therefore, assume that $F(x_1, x_2) \neq 0$ and take $a_1 = a \in \mathbb{R}$ and to $a_2 = F(x_1, x_2)$. We have

$$\begin{aligned} a^2 F(x_1, x_1) &+ a\overline{F(x_1, x_2)} F(x_1, x_2) \\ &+ F(x_1, x_2) a F(x_2, x_1) + F(x_1, x_2)\overline{F(x_1, x_2)} F(x_2, x_2) \geqslant 0, \end{aligned}$$

which amounts to

$$a^2 F(x_1, x_1) + 2a|F(x_1, x_2)| + |F(x_1, x_2)|^2 F(x_2, x_2) \geqslant 0.$$

If $F(x_1, x_1)$ this trinomial in a must be non-negative for every a, which implies

$$|F(x_1, x_2)|^4 - |F(x_1, x_2)|^2 F(x_1, x_1) F(x_2, x_2) \leqslant 0.$$

Since $F(x_1, x_2) \neq 0$, the desired inequality follows.

Theorem

A real-valued function $G : S \times S \longrightarrow \mathbb{R}$ is a positive type function if it is symmetric and

$$\sum_{i=1}^{n} \sum_{i=1}^{n} a_i a_j G(x_i, x_j) \geqslant 0 \tag{2}$$

for $a_1, \ldots, a_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in S$.
In other words G is a positive type function iff $(G(x_i, x_j))$ is a positive-definite matrix for any $x_1, \ldots, x_n \in S$.

Theorem

Let S be a non-empty set. If $K_i : S \times S \longrightarrow \mathbb{C}$ for $i = 1, 2$ are functions of positive type, then their pointwise product $K_1 K_2$ defined by $(K_1 K_2)(x, y) = K_1(x, y) K_2(x, y)$ is of positive type.

Proof

Since K_i is a function of positive type, the matrix

$$A_{n,K_i}(x_1,\ldots,x_n) = (K_i(x_j,x_h))$$

is positive, where $i = 1,2$. Thus, such matrices can be factored as

$$A_{n,K_1}(x_1,\ldots,x_n) = P^{\mathsf{H}}P \text{ and } A_{n,K_2}(x_1,\ldots,x_n) = R^{\mathsf{H}}R$$

for $i = 1,2$. Therefore, we have:

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n} a_i K_1(x_i,x_j)K_2(x_i,x_j)\overline{a_j} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i K(x_i,x_j)\cdot\left(\sum_{m=1}^{n}\overline{r_{mi}}r_{mj}\right)\overline{a_j} \\
&= \sum_{m=1}^{n}\left(\sum_{i=1}^{n} a_i\overline{r_{mi}}\right)K(x_i,x_j)\left(\sum_{j=1}^{n} r_{jm}\overline{a_j}\right) \geqslant 0,
\end{aligned}$$

Theorem

Let S be a non-empty set. The set of functions of positive type is closed with respect to multiplication with non-negative scalars and with respect to addition.

- A function $K : S \times S \longrightarrow \mathbb{C}$ defined by $K(s,t) = (f(s), f(t))$, where $f : S \longrightarrow H$ is of positive type, where H is a Hilbert space.
- The reverse is also true:
  If K is of positive type a special Hilbert space exists such that K can be expressed as an inner product on this space (Aronszajn's Theorem).
- This fact is essential for data kernelization that, in turn, is essential for support vector machines.

Theorem

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

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

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

Which of the following functions are kernels?
For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (x_i + y_i)$$

K is not a kernel. Indeed, for $\mathbf{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$ we have
$k_{11} = K(\mathbf{x}, \mathbf{x}) = 2$, $k_{12} = K(\mathbf{x}, \mathbf{y}) = 3 = k_{21}$, and
$k_{22} = K(\mathbf{y}, \mathbf{y}) = 4$.
The matrix of K is

$$\begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ 3 & 4 \end{pmatrix}.$$

Its characteristic polynomial is

$$\det \begin{pmatrix} 2-\lambda & 3 \\ 3 & 4-\lambda \end{pmatrix} = \lambda^2 - 6\lambda - 1.$$

and has a negative eigenvalue.

$$K_2(\mathbf{x}, \mathbf{y}) = \prod_{j=1}^{n} h\left(\frac{x_i - c}{a}\right) h\left(\frac{y_i - c}{a}\right),$$

where $h(x) = cos(1.75x)e^{-\frac{x^2}{2}}$.
K_2 is a kernel because it can be written as a product $K_2 = f(\mathbf{x})f(\mathbf{y})$.

$$K_3(\mathbf{x}, \mathbf{y}) = -\frac{(\mathbf{x}, \mathbf{y})}{\| \mathbf{x} \| \| \mathbf{y} \|}$$

K_3 is not a kernel because it has negative eigenvalues.

$$K_4(\mathbf{x}, \mathbf{y}) = \sqrt{\| \mathbf{x} - \mathbf{y} \|^2 + 1}$$

K_4 is not a kernel. Indeed, for $\mathbf{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ the matrix

$$\begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix} = \begin{pmatrix} 1 & 5 \\ 5 & 1 \end{pmatrix}$$

has a negative eigenvalue.

Example

A special case of functions of positive type on $\mathbb{R}^n$ are obtained by defining $K : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ as $K_f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x} - \mathbf{y})$, where $f : \mathbb{R}^n \longrightarrow \mathbb{C}$ is a continuous function on $\mathbb{R}^n$. K is translation invariant and is designated as a *stationary kernel*.

Definition

A continuous linear operator $h : H \longrightarrow H$ on a Hilbert space H is positive if $(h(x), x)) \geqslant 0$ for every $x \in H$.
h is positive definite if it is positive and invertible.

If h is an operator on a space of functions and $h(f)$ is the function defined as $h(f)(x) = \int K(x, y)f(y)\, dy$, then we say that K is the kernel of h.

Theorem

(Mercer's Theorem) *Let* $K : [0,1] \times [0,1] \longrightarrow \mathbb{R}$ *be a function continuous in both variables that is the kernel of a positive operator* h *on* $L^2([0,1])$. *If the eigenfunctions of* h *are* $\phi_1, \phi_2, \ldots$ *and they correspond to the eigenvalues* $\mu_1, \mu_2, \ldots$, *respectively then we have:*

$$K(x,y) = \sum_{j=1}^{\infty} \mu_j \phi_j(x) \overline{\phi_j(y)},$$

where the series $\sum_{j=1}^{\infty} \mu_j \phi_j(x) \overline{\phi_j(y)}$ *converges uniformly and absolutely to* $K(x,y)$.

From the equality for the kernel of a positive operator

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

with $a_n > 0$ we can constract a mapping Φ into a feature space (in this case the potentially infinite ℓ_2) as

$$\Phi(u) = \sum_{n=0}^{\infty} \sqrt{a_n} \phi_n(u).$$

Example

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

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

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

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

Example (cont'd)

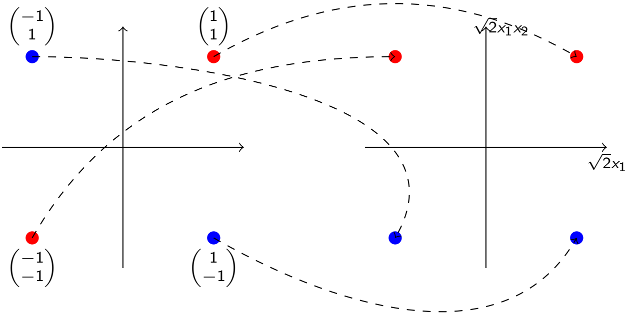
Feature space is $\mathbb{R}^6$

$$K(\mathbf{u},\mathbf{v}) = \begin{pmatrix} u_1^2 \\ u_2^2 \\ \sqrt{2}u_1u_2 \\ \sqrt{2c}u_1 \\ \sqrt{2c}u_2 \\ c \end{pmatrix}' \begin{pmatrix} v_1^2 \\ v_2^2 \\ \sqrt{2}v_1v_2 \\ \sqrt{2c}v_1 \\ \sqrt{2c}v_2 \\ c \end{pmatrix} = \Phi(\mathbf{u})'\Phi(\mathbf{v}) \text{ and } \Phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2c}x_1 \\ \sqrt{2c}x_2 \\ c \end{pmatrix}$$

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 $K(\mathbf{u}, \mathbf{v}) = (\mathbf{u}'\mathbf{v} + 1)^2$ we have

$$\Phi \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ 1 \end{pmatrix}.$$



For the kernel $K(\mathbf{u},\mathbf{v}) = (\mathbf{u}'\mathbf{v}+1)^2$ 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.

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 $K'(u, u) = 1$.

Theorem

Let K be a positive type 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

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

$\mathbf{K}$ is positive, 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,

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

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

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

Theorem

Let K be a positive type kernel. Its normalized kernel is a positive type kernel.

Proof: Let $\{x_1, \ldots, x_m\} \subseteq \mathcal{X}$ and $\mathbf{c} \in \mathbb{R}^m$. 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

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

where Φ is the feature mapping associated to K.

Example

Let K be the kernel

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

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

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

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

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

We prove that K is of positive type by showing that $K(\mathbf{x}, \mathbf{y}) = (\phi(\mathbf{x}), \phi(\mathbf{y}))$, where $\phi : \mathbb{R}^k \longrightarrow \ell^2(\mathbb{R})$. Note that for this example ϕ ranges over an infinite-dimensional space.

We have

$$
\begin{aligned}
K(\mathbf{x},\mathbf{y}) &= e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}} \\
&= e^{-\frac{\|\mathbf{x}\|^2+\|\mathbf{y}\|^2-2(\mathbf{x},\mathbf{y})}{2\sigma^2}} \\
&= e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}} \cdot e^{-\frac{\|\mathbf{y}\|^2}{2\sigma^2}} \cdot e^{\frac{(\mathbf{x},\mathbf{y})}{\sigma^2}}
\end{aligned}
$$

Taking into account that

$$e^{\frac{(\mathbf{x},\mathbf{y})}{\sigma^2}} = \sum_{j=0}^{\infty} \frac{1}{j!} \frac{(\mathbf{x},\mathbf{y})^j}{\sigma^{2j}}$$

we can write

$$\begin{aligned} e^{\frac{(\mathbf{x},\mathbf{y})}{\sigma^2}} \cdot e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}} \cdot e^{-\frac{\|\mathbf{y}\|^2}{2\sigma^2}} &= \sum_{j=0}^{\infty} \frac{(\mathbf{x},\mathbf{y})^j}{j!\sigma^{2j}} e^{-\frac{\|\mathbf{x}\|^2}{2\sigma^2}} \cdot e^{-\frac{\|\mathbf{y}\|^2}{2\sigma^2}} \\ &= \sum_{j=0}^{\infty} \left(\frac{e^{-\frac{\|\mathbf{x}\|^2}{2j\sigma^2}}}{\sigma\sqrt{j!}^{\frac{1}{j}}} \frac{e^{-\frac{\|\mathbf{y}\|^2}{2j\sigma^2}}}{\sigma\sqrt{j!}^{\frac{1}{j}}} (\mathbf{x},\mathbf{y}) \right)^j = (\phi(\mathbf{x}), \phi(\mathbf{y})), \end{aligned}$$

where

$$\phi(\mathbf{x}) = \left(\ldots, \frac{e^{-\frac{\|\mathbf{x}\|^2}{2j\sigma^2}}}{\sigma^j\sqrt{j!}^{\frac{1}{j}}} \binom{j}{n_1, \ldots, n_k}^{\frac{1}{2}} x_1^{n_1} \cdots x_k^{n_k}, \ldots \right).$$

j varies in $\mathbb{N}$ and $n_1 + \cdots + n_k = j$.

Example

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

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

With $a, b \geqslant 0$ the kernel is of positive type.